Quality control for data-dependent and data-independent mass spectrometry-based proteomics

Marina Kriek

Dissertation presented for the degree; Doctor of Philosophy in the faculty of Medicine and Health sciences at Stellenbosch University

Supervisor: Prof David Lee Tabb

Co-Supervisor: Dr. Stoyan Hristov Stoychev

December 2020

Declaration

This dissertation contains work adapted from a publication in the Wiley Online Journal, *Proteomics*, where the co-authors' writings were removed, with the exception of the work of Prof David Tabb who performed the MSGF+ search as well as the Sciex file conversion. The corresponding materials and methods sections are his account of procedure used. The software, SwaMe, was created along with collaborators from the University of Manchester. The parser, Yamato, had been created by Paul Brack previous to the creation of SwaMe and was adapted by him and myself to suit the metrics that I had written. Of the github code available, the section, SwaMe and SwaMe.Test, contains purely my own work, edited and reviewed by Paul Brack and Peter Crowther of Manchester University. SwaMe.Pipeline and SwaMe.Console were created by myself and Paul Brack, later edited by Peter Crowther. MzQCGenerator was originally created by myself, specifically to the SwaMe module, but has been edited and improved by Paul Brack to a very great extent to fit all of the software's needs. All other modules were the sole creation of Paul Brack, edited by Peter Crowther and in some cases, myself.

December 2020

Abstract

Discovery proteomics is advancing at a rapid rate, and quality control of the technique must adapt accordingly. In 2012, a console application, QuaMeter, was created to produce quality control metrics for data-dependent proteomics based on metrics first designed by the USA National Institute for Standards and Technology (NIST). In 2014, the tool gained an identification-independent mode, which can generate 44 quality metrics still applicable only to data-dependent acquisition. However, the development of new data-independent acquisition methods in recent years introduces the need for a data-independent acquisition version of QuaMeter. The QuaMeter metrics must also still be analysed in a statistical framework such as R/Python to gain full value of the multivariate nature of the metrics. Biologists who are inexperienced at programming/ using a console might therefore find the use of such software limiting and there is a desire for a tool with a user interface with which to analyse the metrics.

Here, I have created a console software for the analysis of data-independent acquisition results. The tool provides a platform for in-depth analysis of data quality. The tool is the first of its sort to allow the user to divide the retention time into segments and return quality metrics for each segment separately. This allows the researcher to gain extra insight into the chromatography steps, and as I illustrate here, the tool illuminates problems that would not have been visible if only one metric was provided for the entire file. In addition, the m/z axis is split into the data's underlying isolation window structure and metrics calculated for each window separately to equip a researcher with additional information for method development. A set of metrics is also added which produce one value for the entire file for easy outlier detection among files.

This project also involves the creation of a desktop application with user interface for running either of the two console applications. This tool can also perform some of the key downstream analysis regularly performed in quality control. Outlier detection is enabled via PCA,

classification of longitudinal data as good or bad quality is performed with random forest analysis and individual metrics can also be visualized against their distributions. In addition, many quality control principles are explained and demonstrated in the context of the quality control metrics, such as experimental design, identifying sources of variability in an experiment and conventional quality control techniques such as outlier detection and classification of data quality are demonstrated.

Opsomming

Proteïen-massaspektrometrie maak die afgelope dekade baie vinnig vordering en die gehaltebeheer van die tegniek moet derhalwe dienooreenkomstig aangepas word. In 2012 is ’n konsole-toepassing, QuaMeter, vir die voortbrenging van gehaltemetings vir data-afhanklike proteoomanalise geskep. Hierdie weergawe van die toepassing is op ’n toepassing deur die Amerikaanse *National Institute for Standards and Technology* (NIST) gebaseer. In 2014 is ’n identifikasie-onafhanklike weergawe van die sagteware bygevoeg, wat 44 gehaltemetings rapporteer, maar steeds net vir data-afhanklike verkrygingstegnieke. Meer onlangs is daar egter nuwe data-onafhanklike verkrygingsmetodes ontwerp wat redelike steun in die gemeenskap geniet. Daar het dus ’n behoefte aan ’n data-onafhanklike weergawe van QuaMeter ontstaan.

Die resultate van QuaMeter moet egter steeds stroomaf deur ’n statistiese raamwerk soos R/Python geanaliseer word om die meerveranderlike aard van QuaMeter ten volle te benut. Bioloë wat onervare in programmering of die gebruik van ’n konsole is, mag dit dalk as ’n onoorkomelike struikelblok beskou. Ek het derhalwe ’n konsole-sagteware, SwaMe, vir die analise van data-onafhanklike verkrygingsresultate gebou. SwaMe verskaf ’n platform vir ’n meer diepgaande analise van die datagehalte. Dié hulpmiddel is die eerste in sy soort wat die gebruiker toelaat om die retensietyd in segmente te verdeel en gehaltemetings vir elke segment afsonderlik te bereken. Sodoende kan die navorser insig verkry in die chromatografie, en soos ek hier aantoon, word instrumentele probleme uitgewys wat nie sigbaar sou gewees het indien daar slegs een waarde per monster gerapporteer was nie. Die m/z-as word in die data se onderliggende isolasievensterstruktuur onderverdeel en gemiddelde metings word vir elke venster afsonderlik verskaf, wat metode-ontwikkeling verder vergemaklik. ’n Stel metings wat slegs een waarde per monster bereken, word ook verskaf, wat veral in uitskieteropsporing nuttig is.

Die projek sluit ook die ontwerp in van ’n grafiesekoppelvlak-toepassing, Assurance, wat ’n platform bied om die twee konsole-toepassings aan te wend. Dié werktuig kan ook help met die uitvoering van sekere van die belangrikste stroomaf statistiese analise. Dit word gereeld in gehaltebeheer uitgevoer en sluit in uitskieter-identifisering van hoofkomponentanalise en die klassifisering van longitudinale data as goed of sleg deur masjienleer; die visualisering van individuele metings met die dataverspreiding kan ook plaasvind. Talle gehaltebeheerbeginsels, soos eksperimentele ontwerp en die identifisering van bronne van veranderlikheid, word ook verduidelik en in die konteks van die gehaltemetings gedemonstreer. Daarbenewens word tradisionele gehaltebeheertegnieke soos dataklassifisering ook gedemonstreer.

Acknowledgments

I would like to acknowledge everyone who assisted and supported me during this PhD.

I would specifically like to thank my parents, Kallie and Trix for their encouragement and support. My completion of this degree also coincided with the planning of my wedding, hence the stress that I put these amazing humans through exceeded even what is usually asked of the parent of a PhD candidate. Their trust and love seems endless and I am extremely grateful to have them as role models.

I appreciate my new family and especially Servaas, whom I know will find this paragraph too soppy for his taste. Applying for a PhD degree was part of my list of actions to be completed when I win the lottery and can therefore afford to leave a full-time salary, however, he encouraged me to continue contacting universities and to find a way to accomplish my dream. Throughout the degree he was there to encourage and advise and has truly made me better in what I do.

I would like to thank Prof Dave Tabb for taking this chance on me. I truly loved this project and I feel that I had hit the supervisor and project jackpot. You truly went out of your way to ensure that we receive the best possible education and verified that we were mentally coping throughout, for which I am eternally grateful.

I am thankful for the bioinformatics group in general and specifically Prof Gerard Tromp and Prof Helena Kuivaniemi. You made our group feel like a family where we learnt not only about science specific to our projects, but also about life, birds and snakes. I heard for the first time that it is possible to not only survive, but even conduct wedding celebrations at -40 degrees Celcius and along with the rest of the group critically examined just how strange and curly “boerewors” is when you think about it.

I would like to thank Dr. Stoyan Stoychev and his lab. I enjoyed working with you and being an ad-hoc part of your lab. Thank you for all the advice, comments and for making me feel

welcome up in Pretoria. Thank you also to Dr. Ireshyn Govender who even when I messed up big time did not even yell at me, but merely offered to help and offered advice throughout.

I appreciate all the help, advice and collaborative efforts that spurted from my very fortuitous conversation with Paul Brack at the HUPO-PSI 2019 meeting. It was a true pleasure to work with you, Peter Crowther and the rest of the team. I learnt so much from you and truly enjoyed working with you all. You're a great group of people.

I would also like to thank all of my other co-authors who have all assisted me in my questions, given me advice on science and life.

Of course, I would also like to acknowledge my funders. This work is based on research supported in part by the National research foundation of South Africa, grant number 123237. The first two years of my project were supported by the Council for Scientific and Industrial Research (CSIR).

Trademarked terms

SWATH	Trademarked by AB SCIEX PTE. LTD.
Random Forest	Trademarked by MINITAB, LLC
Spectronaut	Trademarked by BIOGNOSYS, AG

Research output

A research article was published titled: Interrogating fractionation and other sources of variability in shotgun proteomes using quality metrics.
Authors: **Marina Kriek**, Tandeka, U. Magcwebeba, Koena Monyai, Nelita Du Plessis, Stoyan H. Stoychev, David L. Tabb
Proteomics
DOI:10.1002/pmic.201900382

Table of Contents

List of abbreviations

RT	Retention time
LC	Liquid chromatography
MS	Mass spectrometry
LC-MS/MS	Liquid chromatography tandem mass spectrometry
PCA	Principal component analysis
FWHM	Full width at half of the peak maximum
SWATH	Sequential window of all theoretical spectra
DIA	Data-independent acquisition
DDA	Data-dependent acquisition
QC	Quality control
QA	Quality assurance
HUPO	Human Proteome Organisation
HUPO-PSI	Proteomics Standards Initiative of the Human Proteome Organisation
NIST	National Institute of Standards and Technology
USP	United States Pharmacopeia
EuBiC	European Bioinformatics Consortium
FDA	United States food and drug association
SDS	Sodium dodecyl sulphate
PAGE	Polyacrylamide gel electrophoresis
SEC	Size-exclusion chromatography
TOF	Time-of-flight detector
CID	Collision-induced dissociation
HCD	Higher-energy collisional dissociation
MDSC	Myeloid-derived suppressor cells
FFPE	Formalin-fixed paraffin-embedded
PEG	Polyethylene glycol
SILAC	Stable isotope labelling by amino acids in cell culture
TAILS	Terminal amine isotopic labeling of substrates
iRT	Indexed retention time
TB	Tuberculosis
M.tb	*Mycobacterium tuberculosis*
GUI	Graphical user interface
ANOVA	Analysis of variance
iTRAQ	Isobaric tag for relative and absolute quantitation
PSM	Peptide spectrum matches
TIC	Total ion chromatogram
XIC	Extracted ion chromatogram
AIC	Akaike information criteria
IEF	Isoelectric focusing
ITMS	Ion Trap Mass Spectrometry

FTMS	Fourier Transform Mass Spectrometry
GeLC-MS	SDS-PAGE followed by liquid chromatography mass spectrometry
IQR	Interquartile range
Q1/Q3	The first or third quartile
RAM	Random access memory

List of figures

List of tables

Chapter 1: Introduction to discovery proteomics quality control

1.1 Overview

Computational and technical advances have paved the way for the highly sensitive, efficient discovery proteomics analyses we have available today. Liquid chromatography mass spectrometry (LC-MS/MS) is currently the preferred method for protein identification and quantification for many researchers. Advances such as hybrid mass analysers,[1,2] orbitrap mass analyzers,[3] optimisation of time of flight (TOF) detectors for increased spectral acquisition rates, and multi-dimensional separation techniques[4] have contributed to the sensitivity and resolution power of the instrument. The specificity of mass analyzers allows accurate selection of isolation windows for targeted analyses. In addition, the increased dynamic range provided by technologies such as TOF increase the applicability of the technique to discovery proteomics on complex samples or entire proteomes.

Different acquisition methods broaden the scope of discovery proteomics, such as 'shotgun' proteomics, a data-dependent acquisition (DDA) method for protein identification (the name was first applied by the Yates lab in 1998).[5] The technique starts with a scan of all peptide ions within the dynamic range, MS1. In the next step, peptides are selected from the MS1 scan, for example the top 20 most abundant peptides. One by one, these precursor peptides are selected for fragmentation, and the product ions are detected in an MS2 scan. The selected peptide m/z values are also added to a dynamic exclusion list for a specified time period (for example 30s). As the m/z values of both precursors and products are known, identification of the peptides in this process can be managed via standard database search identification algorithms.[6] The

drawbacks of the method are mostly linked to the stochasticity of the selection process. Between runs of the same sample, the peptides selected may differ quite substantially and of the peptides identified in MS1, as low as 16% are targeted for MS2.[7]

More recently, a data-independent acquisition (DIA) method, also known as Sequential Window of All Theoretical mass spectra (SWATH),[8] was developed where broader, overlapping windows (e.g. 25m/z with 1 m/z overlap) covering the dynamic range are selected in m/z order for fragmentation. The fragments are identified by matching to either a public spectral library or a self-created library, for example one created from DDA data on a similar instrument. Here, the MS1 scan is optional and the lack of a precursor to relate to the product ions results in very complicated bioinformatic analysis to identify peptides. In addition, In contrast to the stochastic sampling process of DDA, identification using DIA should theoretically be more repeatable and reproducible. This technique has been made possible by faster scanning technology. Initially the efficiency of the technique also depended on the spectral libraries available and the additional cost of acquiring DDA runs to create a library was a drawback of this technique. However tools now exist for spectral analysis without a library, such as the software tool, DIA-Umpire.[9] For researchers who prefer creating a library from DDA experiments, a recent paper explored including DDA and DIA in the same run. Negating a separate run for library creation, this would lessen the financial burden. They found their technique compared favourably with DIA-Umpire and was able to identify more protein-groups than library-free analysis.[10]

As proteomics advances, so too must proteomic quality control. Each new step adding complexity and contributing variability should be identified. Proper methods to monitor the relevant steps should be established, whether computationally or via bench practices.

According to Whitney and colleagues,[11] quality assurance (QA) refers to the steps taken before a protocol has begun, spanning the planning of the experiment, identification of the potential sources of variability and identification of the proper guidelines and regulations to be followed.

Quality control (QC) refers to the activities that monitor and correct for, if possible/necessary, the data collection and analysis.[11] As part of quality assurance, it is extremely important that a mass spectrometry researcher be able to predict the possible sources of variability before performing an experiment as well as identify the main contributors toward poor instrument performance as well as data quality. Before the experiment, the knowledge of variability factors can be used in the study design as far as possible. During the experiment, quality control analysis may be able to monitor some of these parameters via bench techniques or QC metric producing software and correct for them where necessary. After the data analysis, these sources of variability should be taken into account when discussing results, and if a batch effect occurred, this should be very clearly conveyed in the discussion and the validity of the results discussed.

1.2 Quality assurance

An argument can be made that QA involves the design of the experiment, the identification of possible problem areas, whilst adhering to or at least taking into account guidelines and recommendations.

1.2.1 Institutions, standards and guidelines for proteomic reproducibility and QC

Increasing reproducibility remains an important goal in proteomics, so much so that in a 2002 meeting of the Human Proteome organization (HUPO) the proteomics standards initiative (HUPO-PSI) was initialized.[12] The initiative focuses on setting standards in terms of file formats, controlled vocabularies and more. There is also a working group for quality control specifically which is currently working on a standard file format in which quality control results can be output, mzQC. America boasts the National Institute of Standards and Technology (NIST-accessible at www.nist.gov), an organization that has produced not only countless articles on quality control, but also guidelines and its own set of quality metrics for discovery proteomics

(also known as shotgun) data.[13] In addition, the European Bioinformatics Community (EuBIC) have set up a project which aims to create guidelines for reproducible mass spectrometry experiments. These guidelines will be published to allow their incorporation into journals and data repositories alike.

A researcher in search of mass spectrometry guidelines might find valuable chromatography metrics in the chromatography section of the United States Pharmacopeia (USP). The USP also contains two mass spectrometry sections: one in the general chapters section[14] as well as one for applications of mass spectrometry as chapter 1736 in the 39th issue. However, the USP treats each part (sample preparation, chromatography and mass spectrometry) separately, with the latter left to manufacturer's instructions.

The United States Food and Drug Association (FDA) on the other hand has more specific guidelines.[15] Some of the requirements include for example that at least one control and one fortified sample be run daily. In addition, blank samples should be run after standards/ fortified controls to ensure carry-over does not take place.

It is also a defining characteristic of almost all accreditation facilities to require extensive date-stamped paperwork to back up any quality decision. It is therefore imperative that whichever QC software is used, a report of some kind is generated that can be stored and presented in an audit.

1.2.2 Experimental design

There are two main experimental design methods to ensure that within a study, one group of data is not exposed to higher levels of variability than another.

The first, detailed in 1927,[16] is blocking. This approach involves identifying the sources of variability that cannot be corrected for or changed, for example an experiment where two different batches of a reagent must be used. The effect of this source of variability on the results

is then confounded with other problem areas. In order to maintain a low impact on the results, blocking factors must be established. These include factors in the study that are not directly involved in the biological questions but that may cause between-sample variability. A good example might be the gender of patients in a drug trial. The blocking factor must then be spread equally among the different blocks. Unfortunately, it is not probable that a researcher will be able to think of every possible source of variability, and even if it were possible the sample size of most studies would complicate creating a block for every possible source.

The second involves randomizing the sample order so as to correct for all sources of variability, not just the known sources. This method prevents bias and can distribute all variability evenly. However, this can also cause bias to a certain extent. In a case-control drug trial for example, pure randomization could, by definition, place all the cases together and all the controls together. An issue such as instrumental drift may then cause an effect that may appear to be of biological significance.

The preferred approach is therefore randomized block design as reviewed by Oberg and Vitek.[17] For the variables that cannot be blocked/have not been identified, a researcher is able to apply randomisation within each block. In this case instrumental drift of the LC-MS/MS instrument may prove a useful example where the samples run directly after calibration of the instrument may show more repeatable results. Consequently, the order of the samples should be randomised to protect against a bias. Randomisation should occur within the assigned blocks. There are various algorithms by which to effect randomisation,[18–20] else a random number table could be used similar to one provided by the National Institute for Standards and Technology (NIST) handbook appendix B.[21] This type of experimental design is more easily achieved in some experiments than others. Due to differences between samples originating from different groups, changes in method/instrument may be required which could wreak havoc

if the groups are randomised. However, it is imperative that randomisation be applied as widely as possible.

1.2.3 Sources of variability in a discovery proteomics experiment

In order to effectively implement a randomized blocking strategy, it is important to identify the potential sources of variability present in the experiment. Here, these nuisance factors have been classified into two main steps - sample preparation and data acquisition.

1.2.3.1 Sample preparation

Sample collection

Although QC should be an integral part of study design, implementation of QC steps should begin with sample collection and if the experiment calls for it, sample creation, growth etc. Whether the collection includes harvesting cells, drawing blood or collecting tissue, this process should be done with the source of variability in mind. Hassis and colleagues in 2015 found the majority of plasma-related sample handling errors occur due to recontamination with whole blood during centrifugation,[22] whereas Geyer and colleagues, 2019, found platelets, erythrocytes and coagulation to be the most common quality concerns for this sample type.[23] Environmental variables such as temperature, humidity, buffer/media composition should be closely monitored. Any researcher who has run an autoclave frequently enough is aware that even the best attempts to create a reproducible run can result in significant colour or volume differences in growth media between autoclave runs, which in turn can affect biological conclusions. Similarly, batches of swabs, petri dishes and other consumables presumed to be sterile can be contaminated before receipt. Batches of sterile consumables, reagents, autoclave runs etc should therefore be kept constant within a block.

In addition, the complexity of the proteome is such that slight sample handling errors can cause statistically significant differences in the measured proteome.[22] Geyer and colleagues, 2019, published recommendations for plasma quality including panels of proteins that can serve as contamination indices, along with a tool with user-interface with which MaxQuant[24] results could be added and analyzed for contaminants. [23] The authors, after doing a meta-analysis found 54% of biomarker panels they analyzed to include what they consider to be contaminants/ preparation artifacts.

It is important to develop a standard method or in the case of more than one sample collector, a standard operating procedure for the sample collection. Different sample collectors and more broadly sample sites should be included in the blocking structure if possible.

Sample composition

Samples containing protease inhibitors or anticoagulants could interfere with digestion. Similarly, samples containing proteases could affect results.[25,26] Guiding bodies such as HUPO have acknowledged sample type inequality and have recommended the use of plasma over serum samples due to increased reproducibility.[27] In addition, the location of proteins in a sample can have an effect. For example, it can be difficult to solubilize membrane-bound proteins.[28]

Sample storage

Although sample type again causes variation in this factor,[26] sample storage eventually affects the viability of all sample types. In addition, fixing samples with formalin and embedding in paraffin (FFPE) can result in two main issues for proteomics. Firstly, a variety of modifications can result from the fixation and secondly, fixation creates reduces the solubility of proteins.[29–31]

Protein extraction

Protein extraction method is highly dependent on various aspects of the study. The sample type may require the removal of artifacts such as paraffin or disintegration of a frozen sample.[31–35] The characteristics of proteins in question may require a different temperature or pH for extraction.[36] The location of the desired proteins within the cell may also play a role. Membrane proteins are a good example as their extraction requires alkaline or acidic buffers, detergents, salts or organic solvents.[37–40] The biological question of the study may therefore require severe changes to be made to the extraction protocol. A study that is looking for all possible proteins present in the entire cell may even require multiple extraction phases.

Protein solubilization

In this step again the biological question plays a role. For example, solubilization under native conditions will have to be performed for protein interaction studies. For such cases mild detergents such as Tween and CHAPS can be applied. On the other hand use of the anionic detergent sodium dodecyl sulfate (SDS) will abolish all biological activity and interactions but this is an efficient approach for unbiased solubilization of the total proteome. Chaotropes such as urea are other popular solubilization alternatives. Urea, however, has been shown to result in carbamylation of the peptide N-termini and Lys/Arg side chains, interfering with digestion. In addition, it also affects ionisation and leads to unexpected retention time (RT).[41] This is further complicated by the age and environmental conditions such as temperature and pH playing a role in the efficiency of the denaturation as well as the degree of methylation.[42,43]

Reduction and alkylation

Disulfide bonds within the protein structure need to be broken in order to completely unfold the protein prior to proteolysis. In order to prevent the thiol groups from forming another disulfide bond, the thiol group is acetylated in a process called alkylation. Over-alkylation, however, can

result in N- as well as S-carbamidomethylation or N- or O-alkylation, which can interfere with identification by changing the expected m/z value of the peptide/fragment.[44]

Digestion

The choice of enzymes for digestion purposes is crucial as the enzymes differ not only in cleavage sites, but also in reproducibility, specificity and sensitivity. Trypsin is the most commonly used enzyme for this purpose due to its high specificity and its production of peptides with a basic C-terminus that are ideally suited to ionisation and fragmentation.[45] In addition, there is a certain safety in using the most popular enzyme, such as reduced cost and widespread availability as well as having increased support amongst bioinformatics toolsets. Between batches, trypsin has been reported to show relative reproducibility, with the origin (bovine or porcine) showing significant effect on the reproducibility of the study.[46] In addition, slight changes in pH during digestion, incubation temperature,[47,48] pH,[48] incubation time with the sample,[49] as well as enzyme-to-substrate ratio can result in differences in efficiency.[50] However, even amongst replicates using an enzyme of the same origin, some variation can be seen.[51] This is of particular concern in a technique such as LC-MS/MS where the volumes in question leave very little margin for error.

Pipetting variation

The small volume also increases the need for an extremely reproducible pipetting procedure. Factors to consider include the handler's technique and the tip shape and material. Even the time spent in the hand of the handler might warm the internal temperature of the pipette, resulting in a slight variation in suction power.[52]

Depletion and enrichment

The presence of highly abundant ions in a sample can mask less abundant ions and prevent their detection, hence depletion of highly abundant proteins or enrichment of less abundant

proteins/peptides is often performed to solve this problem. There are many different methods for performing depletion, and commercial kits and columns available for this purpose vary in reproducibility.[53] However, studies have also shown that depletion can reduce variance in peak ratio, reducing variability in the mass spectrometry process itself[54] or simply reducing CV's of the quantification results.[55] Other studies warn about the importance of carry-over when multiple-use depletion devices are used and recommend, where possible, single-use devices.[56,57] The pH during the enrichment process is of particular importance to maintaining high levels of efficiency.[58] The difficulty in using either depletion or enrichment is to do so reproducibly. Hakimi and colleagues stated that in their lab, enrichment has shown to be more repeatable than depletion, however, a larger sample volume is required.[59]

Fractionation

Another approach to reducing dynamic range and complexity is through the use of various fractionation techniques that result in the subdivision and thus simplification of the source sample, each time only processing a fraction of the peptides. Fractionation can occur at protein or peptide level and some of the more common characteristics on which peptides/proteins are fractionated include size, hydrophobicity and mass. Both size-exclusion chromatography(SEC) and Sodium-dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) has been shown to result in the analysis of proteins that differ from their expected m/z values, possibly due to protein complexes rather than individual proteins being separated in cell lysates.[60] Fractionation techniques may also result in sample losses.[61] In addition, techniques vary in their reproducibility,[62] and the use of chemicals such as SDS can cause significant problems later in the process.

Sample cleanup

All of the above-mentioned steps will require the addition of chemical reagents and the interaction of these with each other as well as with the column and mass spectrometer must be very carefully considered. For example, SDS is known to inhibit digestion and will also result in ion suppression as would presence of other detergents and polymers such as polyethylene glycol (PEG).[63,64] The effect of SDS on digestion is also inconsistent among proteins, adding to variability.[65] Salts in the sample can interfere with the ionisation procedure, thereby affecting end results. Sample cleanup methods, which could be based on solid-phase extraction (SPE), precipitation or ultrafiltration are often multistep and thus differ in efficiency, further contributing variability to the experiment.[66]

Labelling

A researcher may choose to label samples in a comparison study at either protein or peptide level. Stable isotope labelling strategies can loosely be classified as belonging to either isotope coded affinity tags,[67] isobaric tags,[68] N-terminal tags, stable isotope labeling with amino acids in cell culture (SILAC),[69] or terminal amine isotopic labelling of substrates (TAILS). Through these various labelling options, multiple samples can be included in the same run and their relative abundances compared, whereby issues of instrumental drift or other intra-run complications are applied to both samples/conditions. However, labelling itself introduces variability into the analysis. Labels have been known to bind to contaminants or simply be subject to incomplete binding.[70]

1.2.3.2 Instrument based variability

In a study by Piehowski and colleagues, the variability contributed by the instrument was less than the variability contributed by sample extraction, but higher than that of the digestion.[51] The

high level of variability contributed by the instrument is problematic, as much of the system is closed and checking the quality mid-analysis is often not an option.

Chromatography

During the chromatography stage, solvent composition can have disastrous effects on the retention time (RT), whilst ionic strength and consistency of mobile phase and composition of buffer, column particle size, length, flow rate and gradient can greatly affect resolution.[71–76] The large influence of detergents and solvents on column efficiency further emphasizes the importance of reproducible and effective sample clean up. The column is also subject to degradation, which manifests as changes in the RT, peak shape or column efficiency/selectivity which may be mistaken for biologically significant results.[77,78]

Ionization

Sample complexity again plays a role in the realm of ionisation due to matrix interference. Inefficiency of desalting and protein extraction could create problems here, resulting in insufficient or selective ionization.[79,80] This highlights the importance of proper sample cleanup. In addition, solvents and electrolytes can also suppress ionization and spray position in relation to the MS opening play a role.[81]

Instrumental platform

Several other instrument problems have been noted and may cause mass error such as power supply voltage changes, temperature/ humidity fluctuations, vacuum system problems.[82] Evaluating the instrument as a whole, the Thermo Fisher Orbitrap has been shown to have increased repeatability in measurements when compared to the LTQ model from the same vendor.[83]

Cycle time

Cycle time refers to the time between the acquisition of the entire set of scans to the next start of the same sequence, for example the time from one MS1 scan to the next. A short cycle time and therefore fast acquisition is imperative for quantitative proteomics.[8] During fragmentation of DDA and most DIA methods, only a small number of ions are selected for detection. Any other ions eluting during an MS2 scan will not be interrogated and are lost to the analysis. It is therefore very important that the time spent to collect a cycle be as short as possible.

Isolation window structure

Different patterns exist for DIA isolation window structure. The classic original pattern described by Gillet and colleagues in 2012[8] involves moving sequentially through the m/z range with windows of around 25m/z with an MS1 scan at the start of each cycle. The MacCoss laboratory has tried a few other methods. A notable example is the MSX method which involves the acquisition of five isolation windows per scan that are a mere 4 m/z in size randomly scattered over the dynamic range.[84] In another example, the group sets an offset for the windows, where the cycle consists of two rounds of windows covering the dynamic range, where the second round is offset by 10m/z.[85] In addition, instead of using windows of a fixed size, variable window sizes have also been suggested.[86] The latter strategy involves software analyzing the peptide distribution of the sample, then allocating smaller windows to the more densely populated sections of the m/z axis. This strategy has been shown to result in a 10-13% increase in the number of identified proteins.[86]

Under-sampling

In DDA, undersampling is the under representation of proteins (usually of low-abundance) in the LC-MS/MS results. Precursor sampling for this acquisition technique is stochastic in nature and although a method like top20 might be used to select the 20 most abundant precursors for

fragmentation, the exact intensity of peptides differ between runs and the chosen peptides therefore also differ. Wang and colleagues make the argument that proteome coverage is directly linked to reproducibility due to undersampling.[63] One study showed as little as 30-65% overlap in identified peptides between two technical replicates.[64]

Trap/detector saturation

Saturation in mass spectrometry is the tendency of a detector or in the case of certain types of mass analysers to not be able to detect ions that exceed a certain abundance. The result is usually a forward-tailed peak and/or a peak with a flat top, flattening at the point of saturation.[87] Intensity is quite variable between runs and exact intensity is almost never compared between runs. By causing the ratio to differ in samples where saturation took place, reproducibility as well as biological conclusion could be greatly affected.

1.3 Quality Control

The identification of possible sources of variability is of course not enough. A proper quality by design approach also includes monitoring and corrections for quality anomalies where possible/necessary.

1.3.1 Bench solutions

In the context of a biomarker trial, Percy and colleagues[21] recommend not performing peptide enrichment or protein depletion to minimize variability and thereby avoid both extra cost and unnecessary loss of proteins from the analysis.

Regarding denaturation concerns, the age of the urea used is extremely important. The same is true of environmental conditions (temperature, pH) during both denaturation and digestion. Maintaining a reliable laboratory temperature is therefore critical.

Digestion variability can be monitored by commercially available digestion indicators, such as those from Thermo Scientific which consists of a non-mammalian recombinant protein.[88] In addition, studies have shown that decreasing the enzyme to substrate ratio decreases the chance of autolysis of trypsin. At a larger substrate to enzyme ratio, studies have also shown that the enzyme becomes more efficient at digesting smaller substrates rather than proteins of interest, so maintaining the correct balance can therefore aid in reducing variability.[89,90]

In order to overcome the crosslinks formed during formalin fixation, techniques exist that include combinations of high temperatures, sonication, high pressure and reagents such as sodium-dodecyl sulphate (SDS) are often added to the buffer to keep proteins in solution.[31,32,35] SDS is also used in solubilizing membrane proteins. However, SDS clusters can dominate the mass spectra, masking signals from any analytes of interest and must therefore be removed quite diligently.[28]

Another commercial solution that has been created to address a variability problem includes attempting to overcome pipetting technique errors with robotics. Sample handling machines such as *Thermo Fisher's KingFisher™ Flex Magnetic Particle Processor, Leap Technologies CTC Analytics PAL® Sample Handler, Tecan Freedom EVO® Automated* are marketed specifically to address sample handling variability. [26]

An interesting solution has been created for correcting labelling biases. In cases where iTRAQ quantitation was used, ratiometric normalization has been utilized as counteraction measure.[91,92] The approach involves adjusting the peptide ratios to center the distribution of peptide ratios on 1. However, this method itself has many concerns over inaccuracy (Reviewed by Christoforou and Lilley[93]).

Arguably the most powerful QC wet lab tools however, are QC samples. Reviewed in detail by Bittremieux and colleagues,[94] samples can consist of single protein samples such as BSA or more complex samples such as HeLA or *E.coli* digests. These samples are typically run

separately in a discovery analysis, but can be included in the sample if necessary. Several studies have also centered around creating their own proteomic spike-in standards for QC and evaluation of methods.[95,96]

Calibration peptides, frequently referred to by the vendor's (Biognosys) designated name of indexed retention time (iRT) peptides, are often used for calibrating the retention time of the sample.[97] This is especially powerful in the case of DIA, where matching RT to that of a library is so critical. In addition, these peptides can also be used to gain information on the quality of the run as their behaviour shows increased stability in comparison to other peptides, and QC metrics on these peptides can therefore be very helpful. The software QuiC™ from Biognosys can be used to assess the quality of the run by evaluating metrics from these peptides and is freely available.[98]

1.3.2 QC software solutions

The quality control of proteomic discovery LC-MS has grown in the last decade to an established field in its own right with analysis pipelines incorporating QC metrics generation as part of their pipeline. Several software packages also exist to generate QC metrics in Shotgun data such as QuaMeter,[99] QuaMeter ID-free,[100] NIST-MS,[13] PTXQC[101], SprayQC,[102] AutoQC within Skyline,[103] Metriculator,[104] OpenMS - KNIME pathway,[105] Downstream analysis of the metrics generated by such software can be tricky due to the highly correlated, not normally distributed, multidimensional data that are produced. Software packages/toolkits that have been created to aid in statistical analysis and visualization of such metrics include SProCoP for generating control charts as part of a longitudinal analysis in the Skyline pipeline,[106] DO-MS for MaxQuant data,[107] MS-stats-QC 2.0 (previous versions were only designed for SRM) which can also form part of the Skyline package,[108] and qc_analysis under the same repository as iMONDB.[109] However, occasionally the metrics producing software also include visualisation and

downstream analysis capabilities such as with rawDiag,[110] iMONDB,[109] SIMPATIQCO,[111] QCloud,[112] pmartR,[113] and statTarget - a tool for signal drift detection and correction.[114]

As so many sources of variability can together impact the measurement drift, it is essential that this important aspect of QC be monitored and corrected if need be. Apart from statTarget, another tool also exists within the Proteowizard library to correct for instrument measurement drift and systemic bias, namely mzRefinery.[82]

Most identification softwares allow the addition of modifications such as carbamylation or carbamidomethylation as variable modifications. Although this method will only allow the verification to take place after data analysis, it remains a powerful tool for verification.

The same availability of QC software cannot be seen where DIA is concerned. One example of such software is QCMAP,[115] which provides certain metrics for longitudinal data such as HeLa samples and performs statistical analyses and visualisations to display graphs such as boxplots. SProCoP generates longitudinal metrics and control charts for SWATH as well,[106] whereas MSStatsQC provides a machine learning approach for the downstream analyses of the metrics produced by certain QC tools such as SProCoP.[108]

Most of the current DIA software are connected to a single pipeline of data analysis or for the more general tools, present very few metrics and very little insight into the data. There is therefore a significant gap for a software similar to QuaMeter, which provides a plethora of metrics in a universal format for further analysis.

The reasons Shotgun QC software is not directly translatable to DIA and why DDA software cannot be used on DIA are all related to the inherent differences between techniques. For example in QuaMeter ID-free, the precursor is used for peak selection and the peak properties such as FWHM are calculated on those peaks. If QuaMeter ID-Free is run on DIA data, the software will attempt to find precursors which do not exist. Another problem is the size of the files and the amount of data in those files. The parsers that were sufficient for Shotgun datasets

were mainly built around reading and storing all data from the file while running, which requires large amounts of Random Access Memory and storage space if run on a DIA file.

1.3.3 Databases

The rising popularity and immense benefits of open science and data sharing has resulted in an increase in datasets uploaded to repositories of late.[116–118] This has led to journal policy changes, where journals (often especially ones with higher impact factors) now require public data submission[119] and the creation of standards for data submission. In 2011, the ProteomeXchange consortium was founded to create a platform via which data submission could be standardised.[117] The consortium currently consists of seven repositories: JPOST, [120] PRIDE, [121] PeptideAtlas,[122] MassIVE,[123] PASSEL,[124] Panorama Public,[125] and iProx.[126] As of 10/08/2020 PRIDE had reported 11187 datasets in their archive, MassIVE reported 10487 datasets, JPOST reported 121 datasets and iProx reported 939 projects, while PeptideAtlas, Panorama Public and PASSEL did not readily report the number of datasets. This wide availability of data enables not only a re-analyses with newer bioinformatic tools/ pipelines, but also an inspection of data to answer different scientific questions than the data was originally intended for. As the number of available datasets increase, the field of proteomics gains a deeper insight into the natural world and non-model organisms become easier to analyze and readily compare data. Similarly, an independent reassessment of results described publicly is made possible, information that can prove invaluable in the analysis of disputed claims. In the case of bioinformatic tool development, it enables the developer to test a novel tool on datasets from a large number of different laboratories, collected under different experimental and instrumental conditions. The implications for quality control (QC) are tremendous, as the quality and reproducibility of datasets do not need to be accepted as the word of the author any longer and can be verified. Hereby, a novel age in quality control begins where standards are set, guidelines are implemented and verification is possible.

1.4 Research goals

It is imperative that researchers in proteomics apply quality assurance and quality control techniques in their experiments. This involves identifying the possible sources of variability, implementing proper experimental design and using QC metric-producing software on data after the run. Users of the DIA method currently do not have an open source program through which to generate in-depth quality metrics. In addition, tools such as QuaMeter are limited in their accessibility by requiring the user to be able to use a statistical language such as R/Python if a free analysis of the metrics is to be performed.

This thesis aims to design a software tool for DIA QC metric production (named SwaMe) as well as a user interface for the downstream basic analysis of both QuaMeter and SwaMe (named Assurance). Furthermore, the use of quality metrics in the traditional sense of detecting underperforming runs, as well as the novel approach of using quality metrics for inspecting sources of variability and illustrating the importance of experimental design is demonstrated.

There are three questions that a researcher should be able to answer about their data with the aid of quality metrics. Firstly, is the platform performing optimally to enable reproducible proteome analysis? Secondly, can the source of the variation be identified with the aid of the software? Thirdly, can the metrics provide documentation to support the biological conclusions drawn from the data?

In Chapter 2, the use of quality metrics to decipher these questions is illustrated using QuaMeter ID-Free. Chapter 3 demonstrates the analysis of these questions with the novel QC metrics producing software for DIA, SwaMe. Chapter 4 illustrates using Assurance, the user interface for running and analyzing QC results for QC analysis and generating reports to fulfill QA guideline requirements on record keeping.

Chapter 2: Interrogating sources of variability in shotgun proteomics using quality metrics

2.1 Introduction

The rising popularity of big data and large-scale analysis places special importance on standardised quality control (QC) pipelines. In this study, the QC pipeline for QuaMeter IDFree[100] was assessed for multiple instrument types, fractionation strategies, and sample types retrospectively using published proteomic datasets from myeloid-derived suppressor cells (MDSCs), cellular exosomes and *Mycobacterium tuberculosis (M. tb)* cell lysates. Spectral counts were then investigated to compare conclusions.

The cellular complexity and cell wall adaptations of *M. tb* cell lysates pose considerable challenges for protein extraction. Due to the pathogenic significance of this species,[127] health researchers have optimized protein analysis by applying fractionation and by synthesising peptides to produce 97% “complete” spectral libraries.[128] The fractionation techniques used in *M. tb* may each contribute to variability.

MDSCs are a heterogeneous population of regulatory immunosuppressive cells that expand during chronic inflammatory conditions such as cancer and infectious diseases such as tuberculosis (TB).[129,130] These cells have therefore been identified as potential targets in immunotherapeutic strategies.[131,132] Although sufficient quantities of MDSCs can be obtained from tissue in mice and blood in humans, subsequent analysis steps can affect the populations of enriched MDSCs; both mechanical and enzymatic purification processes as well as isolation techniques or cryopreservation can introduce variability. Furthermore, MDSCs require gentle

handling and cool temperatures to avoid interfering with measurement of their biological and functional properties. [133–135]

Exosomes have been identified as key role players mediating MDSC intercellular communication and immunosuppressive potency.[136–138] Common MDSC enrichment techniques include ultracentrifugation and precipitation that may contaminate exosome samples with protein aggregates, salts and polymers. These contaminants can influence post-isolation analysis such as mass spectrometry.[139,140]

Many of the existing papers for QC and reproducibility of biological MS have limited their scope to LC-MS/MS experiments, thus missing the variability associated with fractionation, enrichment, and sample limitations. By drawing upon a wide variety of published experiments, this study will detect sources of variability that have been uncharacterized by earlier work. Recognizing their influence on quality is key to making shotgun proteomics more robust for the fight against tuberculosis.

My aim here is to demonstrate the traditional and unconventional use of quality metrics in detecting outliers and in inspecting the data for sources of variability. In addition, I aim to provide insight into and demonstrate the importance of another very important part of quality assurance - experimental design. A researcher aiming to perform a valuable experiment must take quality assurance and control into account in the experiment. This involves being able to identify sources of variability in their own experiment after analysis as well as identifying outliers in data quality.

2.2 Materials and methods

2.2.1 Datasets

This study investigates the fractionation and sources of variability of 11 published LC-MS experiments. Datasets were selected on the basis of connection to *M. tb* or MDSCs and exosomes (see table below).

Table 2.1 – Datasets included in this study

Sample Type	**Investigator**	**Number of single injection experiments resulting in a raw file**	**ProteomeXchange/ PASSEL Reference**
Mycobacterium tuberculosis	Pandey[141]	123 RAW files	PXD010956
Mycobacterium tuberculosis	Tabb[142]	120 RAW files	PXD006843
Mycobacterium tuberculosis	Aebersold[128]	64 RAW files 49 WIFF files	PASS00886 and PASS00656
MDSCs	Escors[143]	6 WIFF files 10 WIFF files 32 WIFF files	PXD001103 and PXD001106 PXD000805*
MDSCs	Ostrand-Rosenberg[144]	6 RAW files	PXD010215

* ("Outgroup": These non-MDSC data were included because the other 16 WIFFs were too few enable PCA.)

MDSCs and exosomes	Schnölzer[145]	51 RAW files	PXD010804
MDSCs and exosomes	Fenselau[137]	60 RAW files	PXD006204
Exosomes	Jimenez[146]	72 RAW files	PXD001487
Exosomes	He[147]	36 RAW files	PXD004779
Exosomes	Liu[148]	54 RAW files	PXD001339
Exosomes	Dobos[149]	48 RAW files 30 RAW files	PXD004062 and PXD010659

2.2.2 Raw data conversion and metric generation

The .raw files from Thermo Fisher instruments were converted to the HUPO-PSI standard format, .mzML,[150] using MSConvert GUI with peak-picking selected.[151] The SCIEX MS Data Converter produced "Protein Pilot" peak lists for SCIEX TripleTOF sets (https://sciex.com/software-support/software-downloads).

By default, the SCIEX MS Data Converter (last updated in 2012) will process only one .wiff file for each call to the software; "AB_SCIEX_MS_Converter *.wiff" does not produce the desired result. We created the following Windows "batch" file to enumerate all WIFF files in the current directory, converting each to mzML using the converter from SCIEX:

```
@echo off
for %%W in (c:\wiff\*.wiff) do (
  echo infile= %%W
  echo outfile=%%~nW.mzML
```

"C:\Program Files (x86)\AB SCIEX\MS Data Converter\AB_SCIEX_MS_Converter.exe" WIFF %%W -proteinpilot MZML %%~nW.mzML /zlib /index)

2.2.3 Configuration for QuaMeter IDFree

The following is an example of a quameter.cfg file for an Orbitrap or TripleTOF instrument that allows the m/z for a given chromatogram to wander back and forth by 0.05m/z:

ChromatogramMzLowerOffset = .05mz

ChromatogramMzUpperOffset = .05mz

Instrument = "Orbi"

MetricsType = "idfree"

OutputFilepath = "metrics.tsv"

2.2.4 From metrics to principal components

The metric tables were analysed using the R statistical environment. The R scripts employed ten libraries to support advanced features. The full R script is available at: https://github.com/marinaPauw/TabbDatasetQuaMeterAnalysis.

The libraries included “MASS”,[152] which boasts a variety of statistical functions, “psych”,[153] which contains functions for psychometric analysis, “lattice”,[154] which boasts tools for creating graphs ,“lme4”,[155] a package centered around linear models ,“car”,[156] which is aimed at users applying regression, “dendextend”,[157] which offers functions for dendrograms and clustering visualisations, “modelr”, which contains functions for modelling, “tidyverse”,[158] an entire universe of R functions and “ggfortify”,[159] a package for data visualisation.

2.2.5 Principal component analysis

Metrics with a variance <1% or a Pearson correlation coefficient >99% were excluded from the study as little novel information would be provided by including both variables in these pairs, so

one of the two is chosen. Robust PCA was conducted to reduce dimensionality and visualize the experiments.

The robust covariance functions cov.rob() and cov2cor() were used to create the covariance matrix that served as input for the PCA. For reproducible analysis, the set.seed() function should be used with the parameter 1234 to increase reproducibility of the cov2cor() function, which includes a randomisation step.

2.2.5.1 Analysis of grouping within principal components

A standard one-way ANOVA does not account for the interaction within the hierarchy of a replicate structure. In this study, a mixed effects linear model was created from the first component in order to conduct nested ANOVA analysis.[160,161]

The fraction number was considered a measurement variable, while biological replicates were represented as a fixed effect and the technical replicates as the random effect. The functions lmer() and lm() were used to create multiple models for mixed effects models and fixed effects models respectively, whereafter the function r.squaredGLMM() was used to view the r-squared values of each of the models in order to determine the best fit. The evaluation process is unique for each dataset, but importantly, increasing the number of effects will almost certainly increase the fit, eventually resulting in overfitting. It is therefore important to critically evaluate the addition of components from a biological and statistical perspective.(See Heinze et al., 2018[162] for review) The formula of the model found to be the best fit for the Tabb dataset was:

$$y = X\beta + Zb + \varepsilon$$

Where:

y is the response variable (Comp.1);

$X$is the fixed effects design matrix

β is the vector of the fixed effects (In the Tabb dataset it was only the *M. tb* strain.)

Z is the random effects design matrix

b is the vector of the random effects (Biological replicate, Technical replicate and date in the Tabb dataset)

ε is the error vector

lmer(Comp.1~Fxn +Sample+(1|Biorep)+(1|Rep)+(1|dates))

The formula of the model chosen for the Fenselau dataset was:

lmer(Comp.1~EC+IC+(1|Bioreps))

This model was used in a nested ANOVA approach similar to Wang 2014,[100] with a probability of less than 0.05 considered significant for the difference between the first component value for groups such clustering occurring by chance. Within the Fenselau dataset, the term EC refers to the sample type (exosomes vs cells) the term IC refers to the biological state (inflamed vs conventional), Bioreps indicate the mouse the sample originated from. Within the Tabb dataset the term Fxn refers to the fraction the sample originates from and Rep refers to the technical replicate, with Biorep referring to samples originating from different original single colonies. The ANOVA was performed for each of the first five principal components in the Tabb dataset and Bonferroni correction was subsequently applied. With five separate ANOVA's analyzed, the p-value considered significant was now 0.05/5 = 0.01.

The sum of Akaike weights is often used to compare the relative importance of variables within a model.[163] Similarly, all model combinations of the different parameters listed in the formulas above were populated. Subsequently, the Akaike's information criterion (AIC) value for all models containing a parameter was calculated using broom.mixed::glance(), added for each parameter and compared to evaluate the parameter influence on the position of the data point in the first principal component.

2.2.6 Factor analysis

Highly correlated metrics were not removed for factor analysis and another robust correlation matrix was generated for all metrics with a variance of >1%. Kaiser-Meyer-Olkin measure for sample adequacy was performed using the KMO() function in the psych package and the analysis was only performed on datasets which scored higher than the cut-off of 0.60. Bartlett's test of sphericity was then performed using the function cortest.bartlett() also in the psych package with a threshold of <0.001. Maximum likelihood was chosen as the factor extraction method and Varimax as the rotation method. The core function fa() was used and the analysis was performed with the robust correlation matrix as input.

2.2.7 Recognizing outliers through distance

Visual inspection with the Elbow method determined the number of principal components/factors to be included in further analysis. The chosen components were used in the calculation of the Euclidean distance matrix via the function dist().[164] The distances between data points were used to calculate relatedness or to identify outliers. A vector was made up of the median distances for each point in the dataset and the interquartile range (IQR) was calculated. If the median distance for a data point exceeds that of Q3 + 3 x IQR, the value was considered an outlier. Tukey's method was chosen rather than Z-score or modified Z-Score due to its relative robustness against outliers and skewed distribution. The MAD_e method shows similar benefits and could also be utilized (For review, see[165]).

2.2.8 Database search and assembly

The Ensembl FASTA for the corresponding species in each dataset was combined with 72 added proteins representing major contaminant proteins. *Mycobacterium tuberculosis* (downloaded 20170518) included 4090 protein sequences. *Homo sapiens* and *Mus musculus*

(both downloaded 20180308 from Ensembl 92) included 107844 and 65542 proteins, respectively.

The MS-GF+ search engine (20170113 release) was employed in all cases,[166] configuring the software to reflect the correct mass analyzer for MS/MS measurement ("TOF", "LowRes", or "HiRes" Instrument settings), precursor mass accuracy of 20 ppm for Orbitrap measurement or 50ppm for TripleTOF, fixed mass shift of carbamidomethylation to reflect the use of iodoacetamide, isobaric tags for PXD001103 and PXD001106, isotopic labelling for PXD001339, and protease specialization for PXD010659 (see Supporting Information for explicit configuration file changes). In all cases, protease specificity was required for only one end of the peptide, and target-decoy analysis was employed universally, with identifications written in mzID format.[167]

2.2.9 Configuration for MS-GF+ for Thermo "Hi-Lo" experiment

The most common database search configuration used in this paper expected MS measurement of peptide ions in an Orbitrap and MS/MS measurement of fragment ions in a quadrupole ion trap. The call to run MS-GF+ looked like this:
java -Xmx8000M -jar /usr/bin/MSGFPlus.20170113/MSGFPlus.jar -s mzMLs -mod Mods.txt -d database.fasta -t 20ppm -m 3 -inst 1 -tda 1 -ntt 1

The Mods.txt file used in almost all cases contained these active lines:

NumMods=2
C2H3N1O1,C,fix,any,Carbamidomethyl
O1,M,opt,any,Oxidation

The most common adaptations to the configuration supported high resolution MS/MS measurement in a TripleTOF (-m 1 -inst 2) or ion trap (-m 1 -inst 0). We always employed

50ppm precursor accuracy for TripleTOF (-t 50ppm) and 20ppm precursor accuracy for Orbi measurement of precursors (-t 20ppm). These were not optimized values, but they should allow for acceptable identification even if the mass calibration of instruments was needed.

The Escors set required iTRAQ configuration, resulting in a change to the command line (-protocol 2) and to the Mods.txt:

H2CS,C,fix,any,Methylthio

304.205360,*,fix,N-term,iTRAQ8plex

304.205360,K,fix,any,iTRAQ8plex

The Liu set incorporated SILAC-labeled amino acids. While the command line required no special options, the Mods.txt file incorporated different lines depending on whether the unlabeled, middle-labeled, or heavy-labeled peptides were being identified from the RAW files:

[middle]

4.025107,K,fix,any,Lys4

6.020129,R,fix,any,13C6

[heavy]

8.014199,K,fix,any,13C6-15N2

10.008269,R,fix,any,13C6-15N4

Finally, identifying peptides in the Asp-N cleaved, biotin-labeled RAW files from Dobos (specifically PXD010659) required changes to the command line for the protease and alterations to the Mods.txt file for the biotin labels:

226.077598,K,opt,any,Biotin

452.245726,K,opt,any,Sulfo-NHS-LC-LC-Biotin

226.077598,*,opt,Prot-N-term,Biotin

452.245726,*,opt,Prot-N-term,Sulfo-NHS-LC-LC-Biotin

IDPicker 3.1 build 18172 imported raw PSMs to produce a single assembly spanning all MDSC sets, another assembly spanning all human exosome sets, and another assembly spanning all *M. tuberculosis* proteomes.[166] Protein parsimony was applied only in the context of the sample class rather than for each study individually. The PSM FDR was set to 0.5%, and the number of spectra required per protein group was increased until the empirical protein FDR fell below 5%. Isobaric reporter ion intensities were not imported into IDPicker for quantitation as isobaric labelling was not the focus of this study.

2.2.10 T-test comparison of two samples

Comparison of the individual metrics between two different sample groups as well as comparison of peptide counts between two samples was conducted using the Welch's t-test using function t.test() in base R. Normality was tested using the Shapiro-Wilk test (shapiro.test() function in base R). Homogeneity of variance was tested using Bartlett's test (bartlett.test() in base R). As these assumptions were not violated, a non-parametric alternative was not used. Bonferroni correction was applied as three t-tests were conducted so the p-value considered significant was now 0.05/3 = 0.0167.

2.3 Results

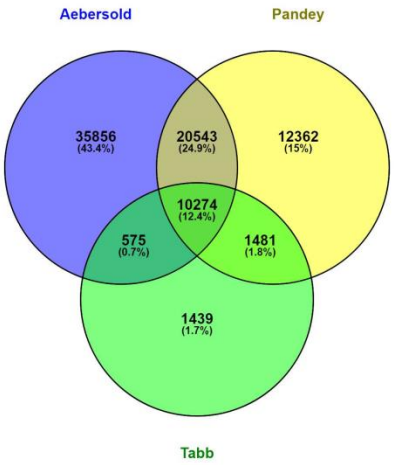
2.3.1 Mycobacterium tuberculosis protein analysis

2.3.1.1 Overlap of distinct peptides from different studies

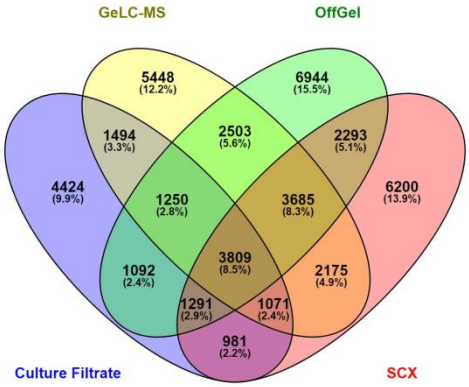
Identified protein counts are often used by researchers as an indication of the quality of a mass-spectrometer run. Here, distinct peptide counts became a focus as they are less sensitive to configuration changes.

The *M. tb* proteome was analyzed from three datasets using a pipeline of MS-GF+ and IDPicker. The relative insensitivity of the Tabb GeLC-MS experiments yielded 13,769 distinct peptides rather than the 67,248 from Aebersold and 44,660 from Pandey. 37.3% of all detected peptide sequences were observed in both Aebersold and Pandey (Fig 2.1). In the case of the dataset originating from the Pandey group, the high levels of sensitivity were achieved by applying four different fractionation strategies and the peptide identification counts are relatively similar between different strategies. Surprisingly, only an 8.5% overlap was found between all four techniques.

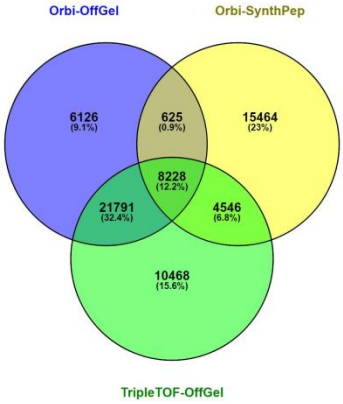
For the Aebersold dataset, OFFGEL was the chosen fractionation technique for both instruments. The TripleTOF showed higher distinct peptide count and a slightly higher unique protein count compared to the Orbi XL, however it should be noted that the OrbiXL is much older technology than the TripleTOF 5500. In addition, the TripleTOF identified 387415 spectra for 45033 distinct peptides, whereas the OrbiXL identified 145693 spectra for 36770 distinct peptides. Both techniques therefore resulted in redundant spectra. The synthetic peptide analysis proved the highest distinct peptide count that was only found in that technique. Contrary to the small overlap seen in the peptide analysis (12.2%), the different techniques applied by the Aebersold study showed 80.3% overlap in the proteins identified. The synthetic peptide method once again led the race with the highest amount of uniquely identified proteins.



A



B



C

Figure 2.1 – Venn diagrams of distinct peptide counts of A) All M. tb datasets (Aebersold, Pandey and Tabb); B) The analysis techniques applied by the Pandey group; C) The different analysis techniques applied by the Aebersold group.

Ensembl lists 4036 protein coding genes for *M. tb* H37Rv. The Aebersold study identified 3873 distinct protein groups, indicating their method development to improve sensitivity was helpful. The Aebersold and Pandey studies together were able to detect 96.7% of the Ensembl listed *M. tb* proteins.

2.3.1.2 Demonstration of outlier detection methods

A critical part of any QC analysis of data is recognizing outlier experiments. Previous studies have employed a supervised,[168] semi-supervised,[169] or unsupervised machine learning approach.[170] In this article, we started with a simple evaluation of distinct peptide counts and compared it to an unsupervised approach, PCA with Euclidean distance computation.[171] We utilized the Tabb dataset for the demonstration.

The dataset contains 120 RAW files from twelve GeLC-MS experiments divided into ten fractions each. Tukey's 3 x IQR criterion recognized three of the fraction 10 runs (SW1-2,SW2-1, SW2-2) and one of the fraction 9 runs(SW2-1) as outliers based solely on distinct peptide counts. As a result of the absence of a block design or randomisation structure, the diminished IDs of fractions 9 and 10 of GeLC-MS SW2-1 are very unlikely to be independent events.

The Euclidean distance matrix created from the PCA of the ID-Free QC metrics[171] highlighted two of the outliers identified by peptide counts (Fig. 2.2) based primarily on MS1-TIC-Change-Q3 and MS1-TIC-Q1:3 metrics which suggests electrospray ionization instability. The internal RAW file dates reveal a 23 day period between these two LC-MS/MS experiments and the next GeLC-MS. Subsequent RAWs did not separate from the rest of the dataset (Fig. 2.3). Evidently, the instrument operator noticed the variation in performance but no runs were repeated.

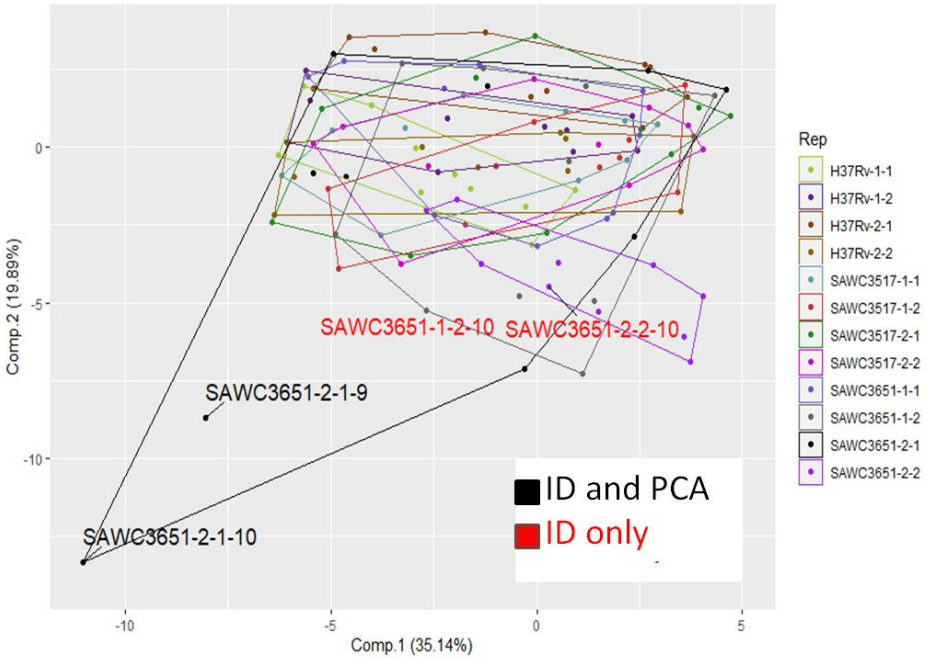


Figure 2.2 – PCA of Tabb dataset. Polygons indicate the GeLC-MS replicates. Two experiments were identified as outliers by identification data only (red) and two more were identified by aPCA of the quality metrics (black). The naming structure indicates M. tb strain (H37Rv, SAWC3517 and SAWC3651), the biological replicate (1,2), the technical replicate (1,2) and the fraction(1-10).

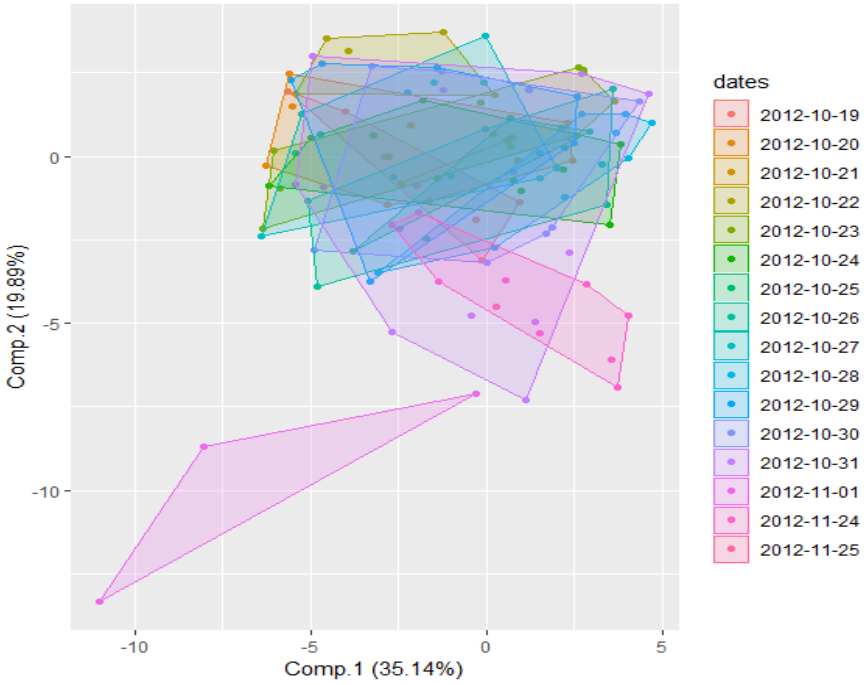


Figure 2.3 – Samples run on different dates grouped together for the Tabb study. Note that the samples run on the last two days return to the larger data group.

2.3.1.3 Investigating the influence of fractionation techniques on variability

Fractionation is known to introduce variability in proteomics,[172–174] and we therefore studied the impact of fractionation on quality metrics. In the Tabb dataset, instead of observing strains or technical replicates grouping together, we observed grouping of similar regions across gels when the first two principal components were visualized (Fig. 2.4).

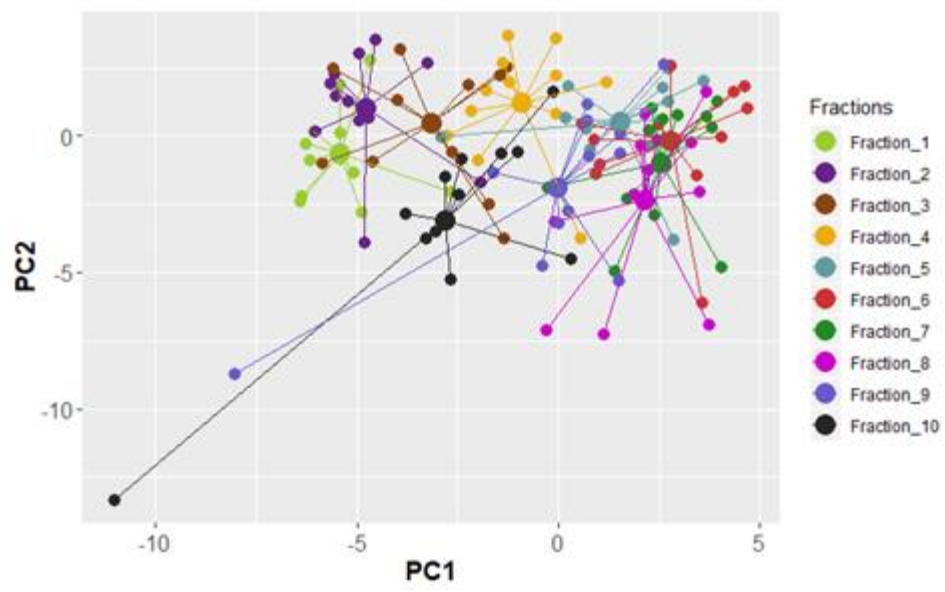


Figure 2.4 - Similar fractions of different samples group together for GeLC-MS dataset in the first two principal components. The mean of each factor group is used as the centroid for each cluster.

Nested ANOVA run for each principal component identified as necessary by the elbow method show the gel region to be the factor with the highest influence on the first four principal components.The p-values from the nested ANOVA were: <2e-16,<2e-16,4.107e-12,2.882e-10 and 0.1373 for component 1:5 respectively. The first four were therefore lower than the 0.01 value considered significant. However, when the sums of AIC weights were calculated, the technical replicate from which a run originates showed a higher relative importance than the fractions. In this case, comparing quality metrics of a fraction to the same fraction in another sample is preferrable above comparing different fractions. If fractionation as a source of variability is ignored, the differences between runs may be misconstrued to be linked to instrument changes.

The nested ANOVA revealed that the distance within the first principal component between each specimen of the same sample type is significantly shorter than the distance between the specimen and specimen of a different sample type with a p-value of 2.734e-08. However, this is only true of the first principal component, not of any of the other four of the first components. To ascertain whether these results would hold up in a different statistical technique, linear models were also created for the PCA results and the sum of AIC weights compared for each variable. When inspecting the sum of AIC weights for each variable, the date on which the experiment was run appeared to play a larger role in the position of each run in the realm of the first principal component than the sample type.

We were surprised to see particular gel regions producing an unusual pattern in the RT.TIC.Qx metrics (Fig. 2.5A). The trend is especially prevalent in SAWC3517, one of the three strains investigated by the study. This set of metrics describes the proportion of all TIC that is accumulated during each quartile of retention time for a given LC-MS/MS experiment. As the four values sum to 1 (all TIC observed in a given RAW), a stacked bar plot of the different quartiles for this metric effectively visualizes the distribution of the TIC. The IDPicker results have been visualized in Fig. 2.5B as a point of comparison. For the region with the highest diversity(later sections of the gel), the majority of signal is located in the third quartile of the retention time. Such information displays the reproducibility in the analyst's technique and could be used to optimize the method for sensitivity by including a higher number of fractions in the highly diverse regions and combining fractions in the less diverse regions.

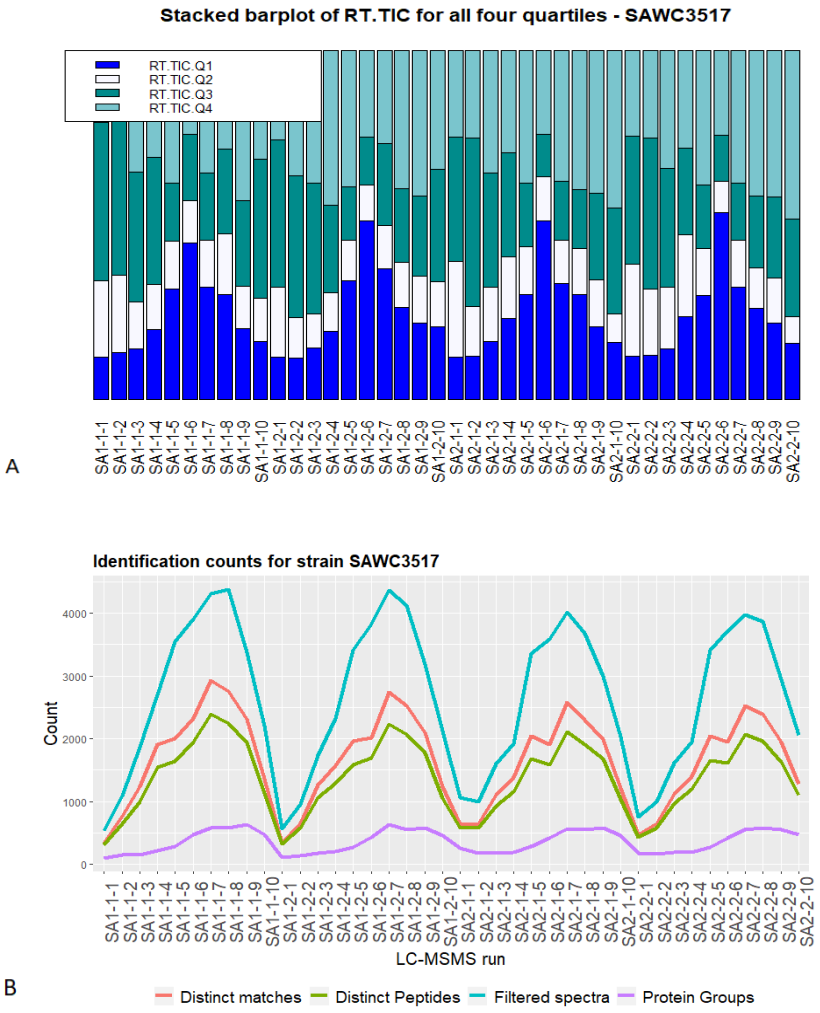


Figure 2.5 A–Similar fractions showed similar RT.TIC patterns in this SDS-PAGE fractionated dataset. The colours indicate the proportion of RT taken to collect each quartile of the TIC. The middle fractions appear to have taken longer to collect the second and third quartile of the TIC. B -Similarly, the number of filtered PSMs, distinct peptides and distinct matches in the data shows a distribution pattern around the different fractions, fraction seven or eight generally showing the highest number of confident peptide spectrum matches (PSMs) of all fractions for a sample.

2.3.1.3 Discerning different experimental parameters and methods

In the Pandey study,[141] a number of different fractionation strategies were employed. Cell lysates were fractionated via GeLC-MS, MudPIT or OFFGEL isoelectric focusing (IEF)(Fig 2.6A) and with experimental settings comprising of either ion trap (ITMS) or Fourier transform (FTMS) combined with collision induced dissociation (CID) or higher energy collision induced dissociation (HCD) (Fig. 6B). The fractionation strategies showed differences in their resulting peak widths (XIC-FWHM-Q3), with cell lysate in gel showing the widest peaks and the Off-Gel strategy generating the highest number of distinct peptides. The data collected via CID-ITMS showed a lower rate of MS2 acquisition and an increased peak width. The Pandey set offers a diversity in MS/MS acquisition strategies that are distinguishable from each other by QC metrics and that overshadow the variability within each strategy.

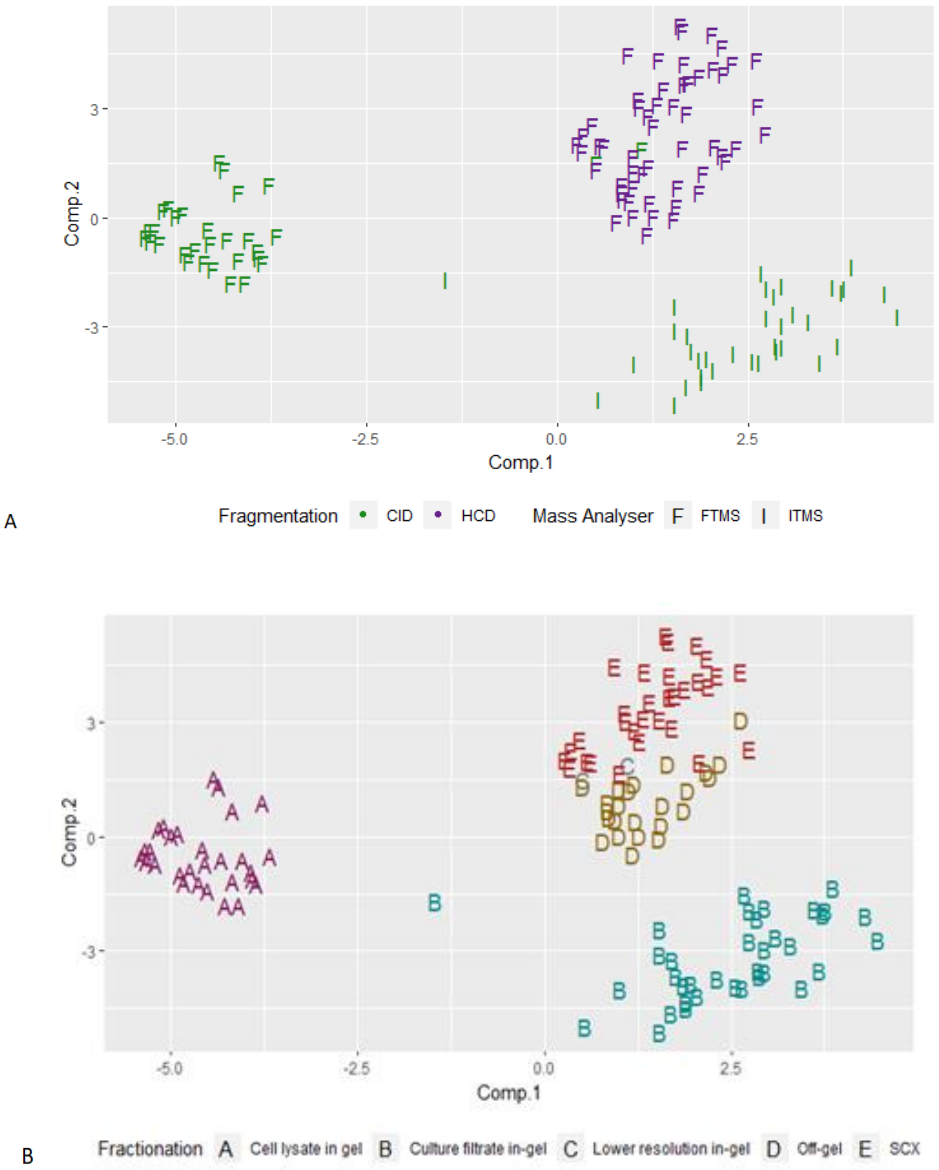


Figure 2.6A – The first two principal components of the Pandey dataset[175] grouped by different instrument parameters(F= FTMS, I = ITMS, green = CID, purple = HCD). B - The first two principal components of the dataset grouped by fractionation strategy.

Similarly, ID-Free metrics are more than sufficient to discern differences between TripleTOF and Orbitrap experiments as well as the difference between peptides originating from cell lysate or synthetic peptides (Fig. 2.7) in the Aebersold dataset.

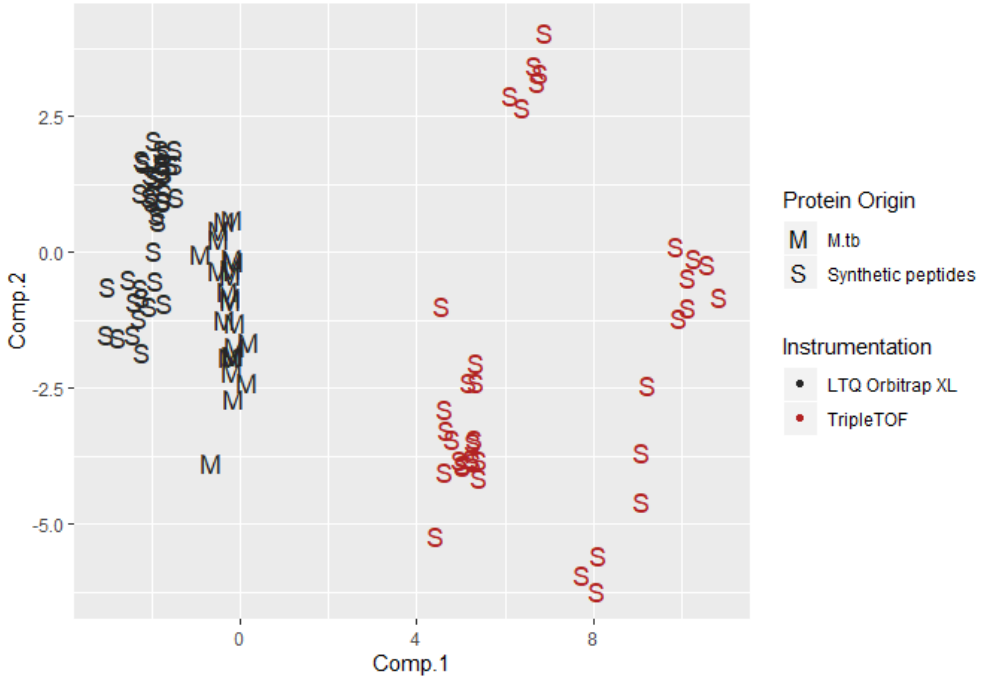


Figure 2.7 - PCA of the Aebersold dataset.[128] Data are grouped by protein origin (S = Synthetic peptides, M = M. tb cell lysate) and the instrument used (colour).

The *M. tb* cell lysates contained fewer peptides than synthetic peptides. The quality metrics of the OrbiXL show higher peptide density (MS2-Density), but lower acquisition speed than the TripleTOF (MS1-Freq.Max and MS2-Freq.Max). T-test results were as follows:

The Welch's t-test results for comparison of MS1 Freq Max and MS2 Freq Max between ThermoFischer Orbitrap and Sciex TripleTOF:

Welch Two Sample t-test

data: metrics$MS1.Freq.Max[1:63] and metrics$MS1.Freq.Max[64:101]

t = -9.945, df = 37.418, p-value = 4.714e-12

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.4032848 -0.9284037

sample estimates:mean of x mean of y

0.942034 2.107878

Welch Two Sample t-test

data: metrics$MS2.Freq.Max[1:63] and metrics$MS2.Freq.Max[64:101]

t = -8.5658, df = 37.576, p-value = 2.299e-10

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-5.199781 -3.211227

sample estimates:mean of x mean of y

2.693013 6.898517

Welch Two Sample t-test

data: metrics$MS2.Density.Q3[1:62] and metrics$MS2.Density.Q3[63:101]

t = 17.479, df = 87.486, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

567.3946 712.9826

sample estimates:mean of x mean of y

792.4194 152.2308

2.3.1.4 Factor analysis

QuaMeter metrics are often highly correlated. PCA assumes that the components created are independent, something that is easily achieved if data is Gaussian, but can often become a problem if variables are highly correlated.[176] Therefore an appropriate alternative dimensionality reduction method to PCA.

When applied to the dataset mentioned above,[142] the visualisation of the first two factors was quite different from when PCA was applied (Fig.2.8A). The clustering of similar fractions was not as strong as with PCA. In addition, FA was unable to detect any anomalies, possibly due to the metrics in which the outliers showed different values to the rest of the dataset, not having been grouped in the same underlying factors (Fig. 2.8B). A researcher must therefore be cautious if only including factor analysis as the main dimensionality reduction method, as it may mask outliers in data quality.

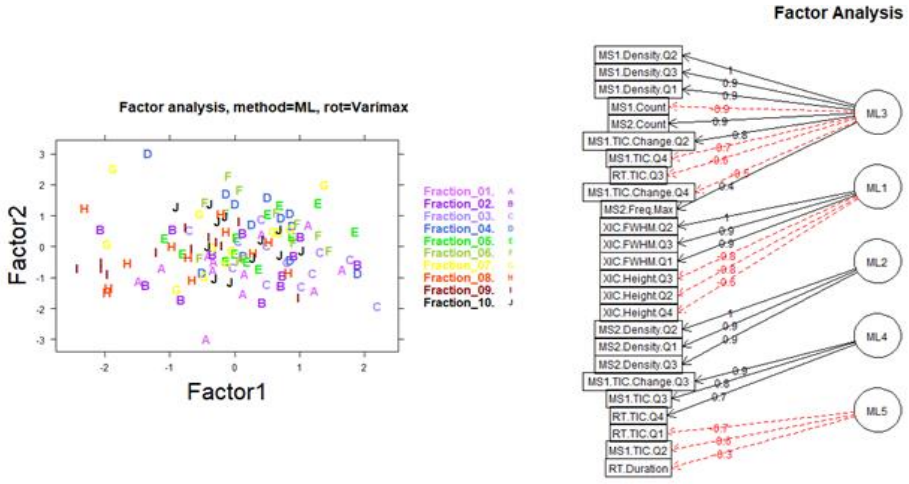


Figure 2.8 – A - Factor analysis of the same dataset which generated Fig. 2. [14] *Note that although some of the similar fractions still group together, the association is much less strong. In addition, outliers are no longer outliers. B – diagram of the first five underlying factors.*

2.3.2 Protein analysis of MDSC's and their exosomes

2.3.2.1 Investigating identification data

The four MDSC experiments we feature enriched for a very small subpopulation of immune cells in murine models. The Ostrand-Rosenberg set was published in 2011,[144] employing older equipment and only six LC-MS/MS experiments. Understandably, 75% of all its identified peptides were also identified in one of the other three sets (Fig. 2.9). The Escors experiment, on the other hand, was the only TripleTOF-powered lab among the MDSC experiments. Approximately half of its peptides were identified only by this team. The anatomical location may also influence the protein content and phenotype,[134] so it is important to note that Escors and Schnölzer experiments obtained MDSC's from bone marrow, whereas Ostrand-Rosenberg and Fenselau datasets obtained MDSC's from blood. Amongst the MDSC datasets, there was not a large discrepancy between Escors, Fenselau and Schnölzer datasets regarding their distinct peptides. The Ostrand-Rosenberg set,[144] conducted long before the rest using older equipment and only six replicates, resulted in only 3.2% uniquely identified distinct peptides and only 0.1% unique proteins identified. The incorporation of technical replicates resulted in a large number of identified spectra for the Fenselau dataset (270270), although these spectra only resulted in 11805 distinct peptides in total and 121 unique proteins. Perhaps the incorporation of fractionation techniques would have resulted in a higher distinct peptide count. Of these datasets, once again the dataset which used a TripleTOF (QTOF) as the instrument resulted in the highest unique distinct peptide count (Escors).

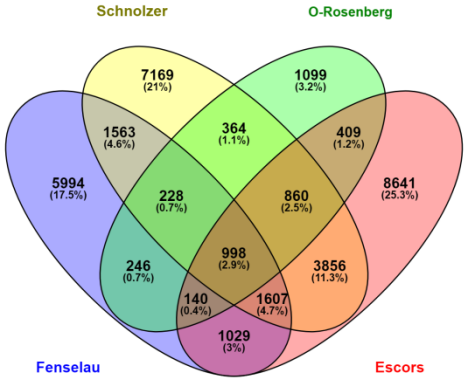


Figure 2.9 - Distinct peptides identified for each MDSC dataset present included in the analysis.

2.3.2.2 Sample type influences QC metrics

When Myeloid derived suppressor cells (MDSCs) are investigated along with their exosomes, QuaMeter ID-Free is able to distinguish the sample types from one another in the realm of the two most important principal components (Fig. 2.10A). It should be noted that in both datasets, the samples were sequentially run and that the dates of the experimental runs follow the same trend, although not exactly separating in the same manner. The second dataset included CD11b+Gr1+ Spleen Cells and MDSCs each with corresponding exosomes. The different sample types once again clustered together, although the separation was not as clear as before (Fig 2.10B).

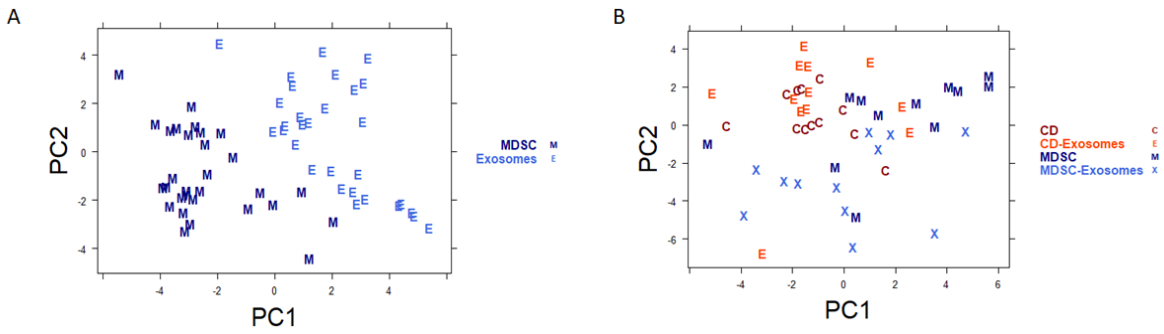


Figure 2.10A - PCA plot showing myeloid derived suppressor cells (MDSCs) and their corresponding exosomes. Note the separation of the two sample types in the realm of the first two principal components. B - In another dataset, the effect was much less pronounced and the separation of samples and their exosomes not as clear.

2.3.3 Exosome protein analysis

2.3.3.1 Peptide overlap in the exosome datasets

The exosome datasets (Fig. 2.11) show great disparity amongst the projects. It must be remembered that the datasets include exosomes originating from different cell types. Of the total number of unique peptides, 77% were identified in the more recent Jimenez dataset.[146] This study incorporated the exosomes of acute myeloid leukaemia blast cells, which generally show a relatively high protein count.

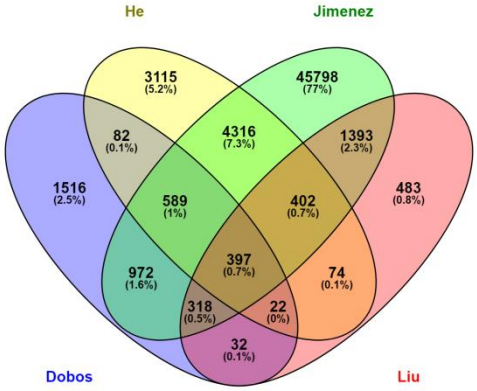


Figure 2.11 - Distinct peptides identified by each exosomes study in the analysis.

2.3.3.2 Fractionation does not always improve sensitivity

In the Liu exosome dataset,[177] samples were either subjected to GeLC-MS/MS or SDS gel direct embedding without electrophoresis. After PCA, the fractionated and unfractionated data separated completely in the realm of the first two principal components regardless of the replicate the sample belonged to, indicating that fractionation played a larger role on the quality metrics than any other step in the method. MS1Density (the number of peaks present in the MS1 scan) was the metric that demonstrated the greatest change; one naturally expects that fractionated samples will show smaller numbers of peptide ions in each MS scan.

Studies often report the number of identified peptides from fractionated data as 21 times,[173] or even 43 times[174] higher than unfractionated data. However, in the Liu dataset, the combination of all fractions of one particular sample did not achieve as high distinct peptide identifications as the unfractionated version of the same sample (Table 2.2). The total distinct peptides identified from fractionated samples and unfractionated samples were not significantly different for this set(analysed via t-test). For Liu, the fractionation strategy yielded limited benefits.

Table 2.2– IDPicker identification data comparing the total of all distinct peptides from all the fractions originating from a single sample, to that of the unfractionated sample.

Group/Source/Spectrum	Distinct Peptides	Distinct Matches	Filtered Spectra	Distinct Charges	Protein Groups
	3519	6292	34641	3	602
Total of Fractionated samples	3080	5579	27311	3	589
Total of all fractions for sample A	1847	2641	8094	3	482
Total of all fractions for sample B	2458	4408	11833	3	554
Total of all fractions for sample C	627	802	2648	3	259
Total of all fractions for sample D	976	1589	4736	3	341
Total of unfractionated samples	1220	1744	7330	3	301
Unfractionated A	927	1247	2055	3	234
Unfractionated B	628	937	1594	3	177
Unfractionated C	764	1096	1894	3	202
Unfractionated D	701	1004	1787	3	188

2.3.3.3 Evaluation of different digestion enzymes

The Dobos dataset included the application of different enzymes. MDSCs were either digested with trypsin, or biotinylated on the lysine residues and digested with Asp-N instead. A 7-month

period separated the two experiments, contributing a substantial batch effect to the QC analysis. 51.1% of the proteins identified in the study were only found in the samples digested with trypsin, with 46.2% identified in both trypsin and Asp-N digests and only 2.6% were uniquely identified in the samples that were digested with Asp-N (Fig. 2.12). Due to its popularity, a large amount of optimization has been performed over the years to give trypsin the advantage. The smaller inventory of proteins from Asp-N in this study could therefore reflect a less-optimized protocol for the rarely used enzyme.

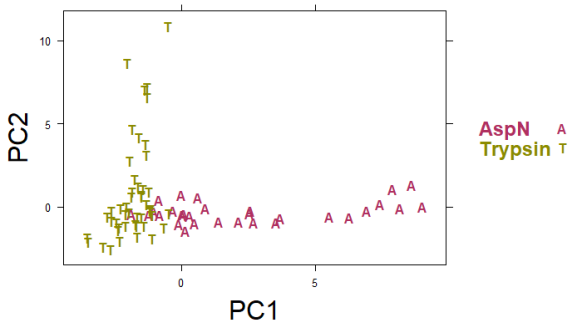


Figure 2.12 – Dataset where two different enzymes were used due to biotinylation of lysine residues. Note that the different enzyme datasets were analyzed with a five-month gap in between (Diaz et al., 2016).

2.3.3.4 Importance of experimental design

Most of the datasets in this study applied a sequential analysis pattern, without blocking or randomisation. However, Jimenez and He datasets did apply a block design to their experimental plan.[146,147] The Jimenez dataset applied SDS-PAGE fractionation to divide samples into 9 fractions. Similar fractions were then run sequentially, starting from fraction 1. This design reduced the impact of any specific event on one biological group (Fig. 2.13).

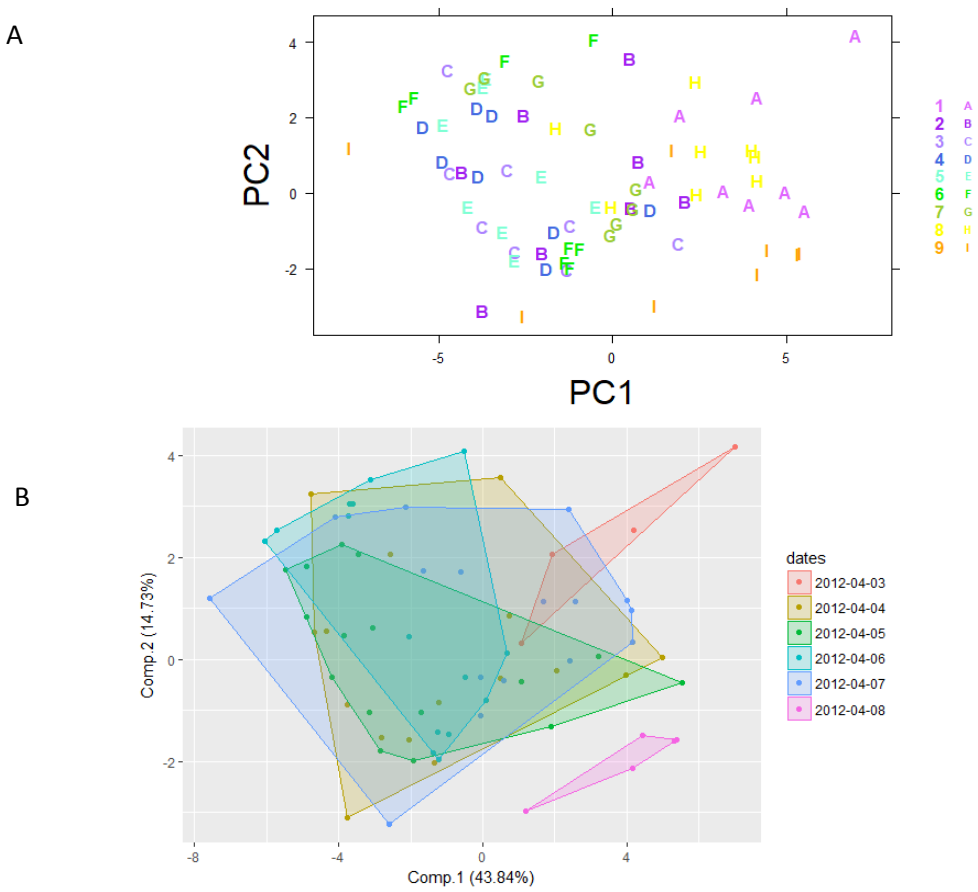


Figure 2.13 – In this GeLC-MS dataset from the Jimenez group, the data was run more or less in a block design, with the same fraction of different biological samples run sequentially. Note that all the runs of a particular date separate from the rest of the experiment (08042018) in B). Those samples correspond to fraction 9 of a number of different samples, therefore the effect of the separation was spread out over different biological samples.

The block design enabled the researchers of the He study to avoid batch effects by not running fractions from the same biological group concurrently. Three samples run sequentially at pH 4 showed a higher MS1.TIC.Change.Q3 than the other three specimens, indicating that for a period in time, the ionization was less stable. However, due to their experimental design, the bias was not applicable to a single biological group (Fig. 2.14 and Fig. 2.15).

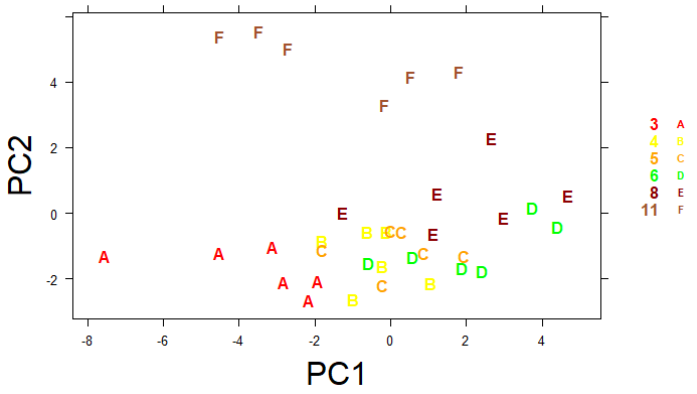


Figure 2.14 – The first two principal components visualized for the He dataset. Note that the data cluster in groups of three, corresponding to the order in which the experiments were performed, as well as to the different pH fractions were eluted at. The samples were run in a block design, avoiding batch effect.

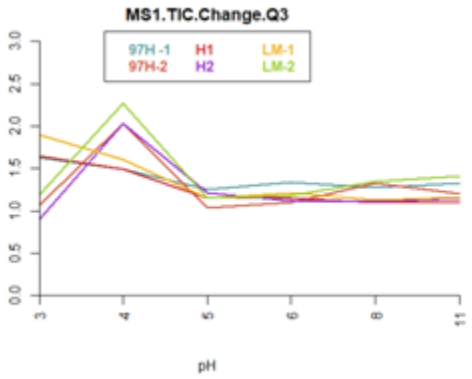


Figure 2.15 – The MS1.TIC.Change.Q3 metric showed increased values for some of the runs, indicating ionization instability. The effect is spread out over the second technical replicate of all three biological replicates. If the block design had not been applied, two replicates of the same group could have experienced ionization instability, allowing incorrect interpretation of the data.

These findings indicate that the three principles for experimental design proposed by Fisher[178] (namely replication, randomisation and block design) help compensate for minor instrumental or technical anomalies within the dataset whilst still producing reproducible results. They illustrate that quality metrics are useful in ascertaining the comparability of fractionated proteomics experiments in an unbiased manner.

2.4 Conclusion

In this study, datasets investigating the complex sample types of *M. tb* and cells and exosomes from the innate immune system of mice or humans were analyzed to illustrate the use of quality metrics to interrogate different stages of the Shotgun workflow. The use of PCA to conduct multivariate analysis demonstrates an under-used avenue for exploring quality metrics. The nested ANOVA approach illustrates the fraction (or SDS-PAGE gel region) of origin for spectra for one of the datasets had a larger influence on quality metrics than the technical or biological replicates the sample originated from. Another dataset saw fractionated and unfractionated samples completely separating in the space of the first two principal components. Fractionation is therefore an integral source of variability that should be considered in experimental design and possibly included in blocking structure. This study has highlighted several parts of an LC-MS/MS experiment as sources of variability, including fractionation technique, sample type, mass analyzer fragmentation method combination, enzyme used for digestion and instrument type.

Lastly, we found that although many of the studies did employ replication, only two studies out of the 11 incorporated blocking/randomisation in their experimental design. These two datasets illustrate that a well-designed experiment can overcome minor technical anomalies that are to be expected in any experiment, while maintaining reproducibility.

Chapter 3 : SwaMe - quality control for DIA mass spectrometry

3.1 Introduction

LC-MS/MS is highly complex, and yet data are not captured for the performance of each aspect of the technology in isolation. The intensities measured for a given m/z at a given retention time must be interrogated closely to detect, for example, a distortion in the liquid chromatography. Software tools used for retrospective QC analysis must therefore provide maximum possible insight to each source of variability.

There are two main acquisition techniques in proteomics. Data-dependent acquisition (DDA)[5] where the full scan (MS1) is used to select a subset of peptides that are subsequently and separately fragmented and analyzed in MS2. Data-independent acquisition (DIA) involves the fragmentation of sequential isolation windows of 3 to 25m/z in size.[8] DDA suffers from an inherently stochastic selection process in the selection of peptides to fragment leading to high number of missing values in particular at the low abundance range where, from a biology perspective, a high percentage of proteins of interest are present. DIA's greatest disadvantage on the other hand could be in dealing with the complexity of this data-intensive technique in order to ensure low levels of false spectrum matches.

Although in-depth quality analysis of DDA metrics is possible via tools such as QuaMeter[99] and QuaMeter ID-Free,[171] the same luxury was not available for DIA analysis. Due to inherent differences between DDA and DIA, using a DDA tool for DIA analysis results in random ions being selected as the precursor peak, which would produce nonsensical results. In addition, DIA

has unique challenges. For example, there are additional factors such as isolation window structure[84,86] that may influence the variability between samples.

The aim of this study is to create a cross-platform, instrument independent, free software that produces QC metrics from DIA data, named SwaMe. I aim to use this software to demonstrate how a researcher may be able to answer two key quality questions about their data.

Firstly, I answer the question whether there are any outliers in the run. Secondly, I answer the question of whether a researcher can determine the main source of quality variability from their experiment results, using SwaMe. Thirdly,, I answer a question that the instrument operator/ quality scientist might ask, namely whether one can be alerted to impending instrumental breakdown by viewing the metrics produced by SwaMe.

3.2 Experimental section

3.2.1 Datasets

Datasets were chosen to represent three different instrument vendors (SCIEX, Thermo Fisher and Waters), as well as a variety of sample types such as cells of bacterial, fungal and human samples, which include urine and tumor tissue. In addition, FFPE tissues were represented as their impact on quality studies has previously been noted.[29] Three of the datasets are publically available with references available (Table 3.1).

Table 3.1 – Datasets included in this study

Sample Type	**Investigator**	**Instrument**	**Number of LC-MS/MS experiments**	**ProteomeXchange/ PASSEL Reference**
Mycobacterium tuberculosis	Aebersold[182]	TripleTOF 6600	36 WIFF files	PASS00655
Homo sapiens FFPE tumor tissue	Hembrough[183]	Q-Exactive	12 Thermo RAW files	PXD010934
Paracoccidioides lutzii	Pereira[184]	Waters Synapt MS	25 Waters RAW files	PXD002285
Homo sapiens urine samples	Stoychev	TripleTOF 6600	150 WIFF files including at least 2 replicates of each file,11 reruns	Data not publicly available
Homo sapiens pediatric urine samples	Steen	Q-Exactive	64 Thermo RAW files(32 trial samples and the same 32 samples rerun	Data not publicly available

3.2.2 File conversion

The .raw files from Thermo Fisher and the .wiff and .wiff.scan files from SCIEX instruments were converted to the HUPO-PSI standard format, .mzML,[150] using ProteoWizard MSConvert command line with peak-picking set to true.[185] For Waters .raw files, MSConvert peak-picking was not employed.

3.2.3 Metric generation

SwaMe is a set of software packages created specifically for the QC of proteomic DIA LC-MS/MS data within the Yamato software framework. This framework consists of multiple parsers. The most important of these enables extremely fast forward-only parsing of an mzML. SwaMe consists of SwaMe.Core and SwaMe.Prognosticator packages that complement each other for the QC analysis of DIA and are combined to produce a single output file for further analysis with a statistical tool. These functions are accessed via a command line interface, SwaMe.Console.

A key design principle of SwaMe is to produce metrics as near to data acquisition time as possible (instead of relying on slow post-hoc analysis), allowing for near real-time analysis of samples. A robust command line interface is intended to automate quality control of samples as they are produced by the LC/MS platform. Metrics are presented as a HUPO-PSI mzQC file,[186] which boasts a JSON format for increased readability and lower memory requirements whilst also storing metadata. A structured schema and controlled vocabulary enables information such as the metric name and description to accompany the value, aiding interpretation and allowing comparison between software tools.

SwaMe was modelled after an ID-Free “shotgun” QC tool, QuaMeter.[171] One key difference is that metrics in SwaMe are separated into three different groups: RT-divided, m/z-divided (by isolation window) and comprehensive metrics. In the RT-divided section, the user is given the

opportunity to provide a value corresponding to the number of segments that the RT is then divided into. Therefore instead of providing one value for the entire RT, the user is presented with a value for each segment of the RT. This section includes metrics that are chromatography-related such as MS2PeakWidth and TailingFactor. There are also general metrics such as the AvgMS2Density and DeltaTICAvg that can provide insight into the distribution of these metrics across the RT.

Each metric in the SWATH-divided section provides a value for every unique target isolation window or “swath.” These include metrics such as the AvgProportionOfTotalTIC which adds the TIC for all swaths of the same target m/z together and then works out the proportion of the total TIC made up by this swath. The swathDensityAverage describes the number of ions on average detected in a swath with the target m/z range in question and similarly the swathDensityIQR calculates the interquartile range of this metric for the particular swath. This section provides information as to the distribution of the metric across the m/z axis.

The comprehensive metrics each provide only one value per run. These metrics include the number of scans such as in MS2Count, the number of ions detected in the MS2 scans in the run such as in totalMS2IonCount as well as a metric reporting the number of scans without any data points.

As part of Addendum A, a full table of the metrics is provided.

For Aebersold,[128] Hembrough,[183] Pereira,[184] and Stoychev datasets, SwaMe was run with the following arguments: tolerance of 0.05m/z, retention time-divided metrics were divided into deciles (divisions=10), minimumIntensity of 100. For Pereira,[184] the minimum intensity was adjusted to 1. For outlier detection in the Stoychev dataset, the divisions were set to 1 and the RT-divided metrics therefore addressed the entire range of RT values rather than portions of

them. The comprehensive metrics and the RT-divided metrics were then combined to form one larger table that could be used as input for a robust PCA.

Downstream analysis was performed in the statistical environment R.

Orthogonal to the core SwaMe metrics, SwaMe.Prognosticator metrics were created by collaborators from University of Manchester. Although SwaMe.Prognosticator forms part of SwaMe, analysis of results can occur independently and as it was created by our collaborators it will not be discussed in this thesis.

3.2.4 Principal component analysis for outlier detection

Each type of metric (metrics provided for each m/z target window, each segment of the RT or metrics provided for the entire experiment) was analyzed separately for each experiment. First, metrics with low variance (<1%) and high correlation (>99%) were excluded as no new information would be provided by such metrics.

For Aebersold, Hembrough, and Pereira, the number of files were too few to enable a robust PCA, and a PCA was performed using the base R function, “prcomp”, with scaling set to true. However, in the Stoychev dataset this method could be conducted similar to what has been described previously.[100]

3.2.5 Identifying outliers through distance

The Elbow method determined the number of principal components/factors required to characterise quality metrics.[163] The chosen components were used in the calculation of the Euclidean distance matrix via the function “dist”. The distances between data points were used to calculate relatedness or to identify outliers. A vector was made up of the median distances for each point in the dataset and the interquartile range (IQR) was calculated. If the median

distance for a data point exceeds that of Q3 + 3 x IQR, the value was considered an outlier (As reported previously[187]).

3.3 Results and Discussion

3.3.1 Introduction

In this study it was important to include publicly available datasets from different instrument vendors to illustrate the versatility of the software toolset and to bring to light the availability of data from different vendors. Although Sciex and Thermo Fisher instrument data forms the majority of data available in the ProteomeXchange[117] repositories, we were pleased to find data from Synapt-MS instruments. The number of MS^E datasets with the keyword “MSE” listed at the time of submission in ProteomeXchange was only 33. PRIDE[121] listed 47 datasets from a Synapt instrument and MSE as keyword. Panorama public[125] only contained one dataset with a Synapt instrument and MassIVE[123] showed 9. The same keywords when input into OmicsDI[188] resulted in 87 matches for PRIDE and only 1 match for MassIVE. Although the percentage that these datasets make up out of the group of scientists that utilize MSE technology is unknown, the scientific community could definitely benefit from an increase in the total number of publicly available datasets from Waters instruments using this technology for acquisition.

With the increased analyses of big data as well as the increased size of DIA files compared to early technologies such as DDA comes the need for faster analysis. Parsing speed was therefore a priority and on a 64-bit i3-3217U laptop with 8GB of RAM the analysis was able to complete in 53s for a standard 3.5GB mzML file from the Stoychev dataset.

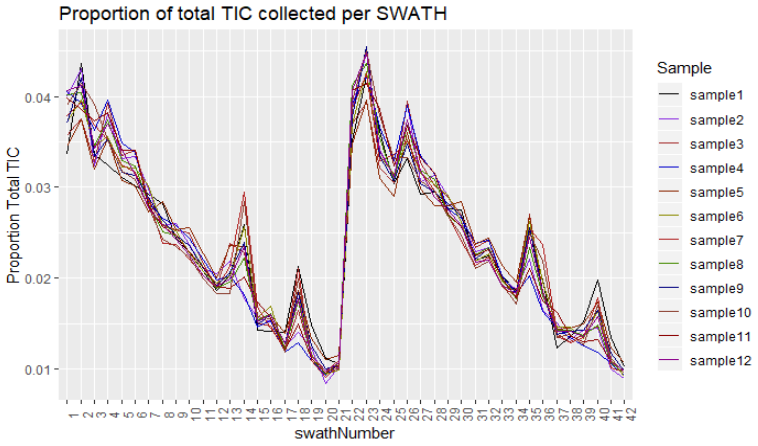
3.3.2 Investigating window isolation scheme with QC metrics

The metrics that are provided per SWATH (therefore metrics are averaged for each window with the same isolation target m/z) provide insights into quality issues that are visible along the m/z

axis. The Aebersold dataset[182] included 32 sequential windows that were a fixed size of 25m/z each. If PCA is performed on the SWATH metrics, it is clear that metrics of the same isolation window cluster together in the space of the first two principal components, indicating the large impact that the m/z of a precursor has on QC metrics (Fig. 3.1A). This behaviour follows what might be expected of a fixed window structure where the ion distribution is not equal between different windows of the same scan. The shape of the PCA results akin to the less than sign can be attributed to the windows at the start of the cycle displaying an increased proportion of total TIC to the middle and end of the cycle, whilst swaths in the middle of the cycle show a higher average density and IQR for the density compared to the swaths at the end and beginning of the cycle. The swaths at the end therefore had lower TIC and lower average density.

In the Hembrough dataset,[183] a staggered isolation window scheme was implemented similar to previous studies.[85,189] Twenty-one fixed, sequential, non-overlapping windows with a width of 20m/z covered the m/z range, only to be followed by the same scheme with each window target m/z offset by 10m/z. This method allows detection of the ions that were at the edges of the windows in the first half of the cycle and has the added benefit of a decreased cycle time compared to conventional overlapping window schemes. After 42 windows had been collected, the cycle was restarted. Despite what is mentioned in the article, the mzML files show no evidence of an MS1 scan being collected. The TIC and average density for the two halves of the cycle show equal average densities and equal proportions of the total TIC. The researchers therefore were able to gain viable peptides from both offsets. SwaMe does not utilize identification, so it is unclear if the number of distinct peptides detected by the two halves of the same isolation window scheme differ. This trend once again shows the strong impact that the m/z of the window has on the QC metrics (Fig. 3.1B).

A



B

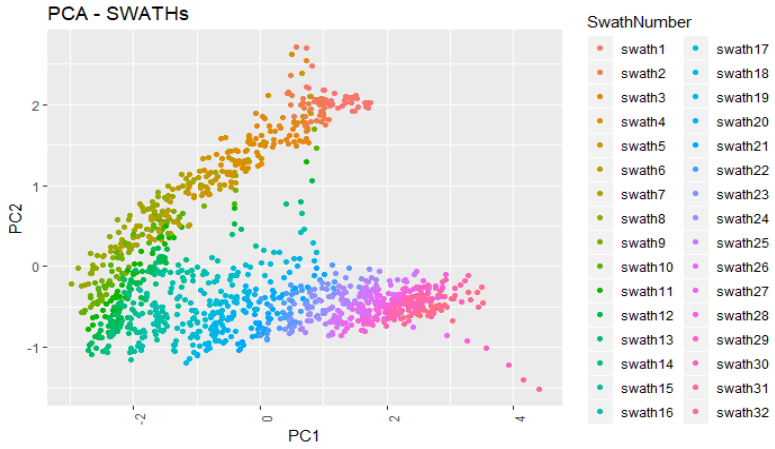


Figure 3.1A - PCA of the swath divided SwaMe metrics of the Aebersold dataset.[128] Each datapoint represents one unique isolation window within the run. B - Proportion of TIC for the entire experiment that was collected by SWATHs of the same target m/z of the Hembrough dataset.[183] Notice the unique isolation scheme of 21 sequential windows followed by the same structure offset by 10m/z to make up 42 unique isolation windows per cycle. No MS1 was collected.

3.3.3 Troubleshooting problematic data with quality metrics

In the Steen dataset, 32 samples were run before a turbo-pump failure prompted researchers to halt the experiment. After replacing the pump, the same samples were rerun. This dataset is perfect for analysis with SwaMe, where the run can not only be analyzed by itself, but also compared with the reruns after.

When viewing the average MS2 density for the different segments of the RT in the Steen trial dataset prior to turbo-pump failure, a pattern emerges of a delayed bell curve with an apex around segments 5-8 (Fig. 3.2A). However, unexpectedly, the average MS1 ion density does not show a clear pattern, with many samples showing a very low density in MS1 scans, others only reaching a density of 500 ions detected after segment 8, while some never dropped below 500. When comparing the trial and rerun values for this metric, we see that this initial decrease in ions detected was unique to the trial run (Fig. 3.3A). In addition, although the second to last run sample did not deviate from the usual trend seen in the rerun, most of the deviations occurred in close temporal proximity to the pump failure (Fig. 3.3B). This trend is not reflected in the total MS1 (Fig. 3.2B) or MS2 TIC (Fig. 3.2C), nor does the change in TIC from one scan to the next (Fig. 3.2D) show much deviation from the total TIC as one would expect with ionization problems such as sputter.

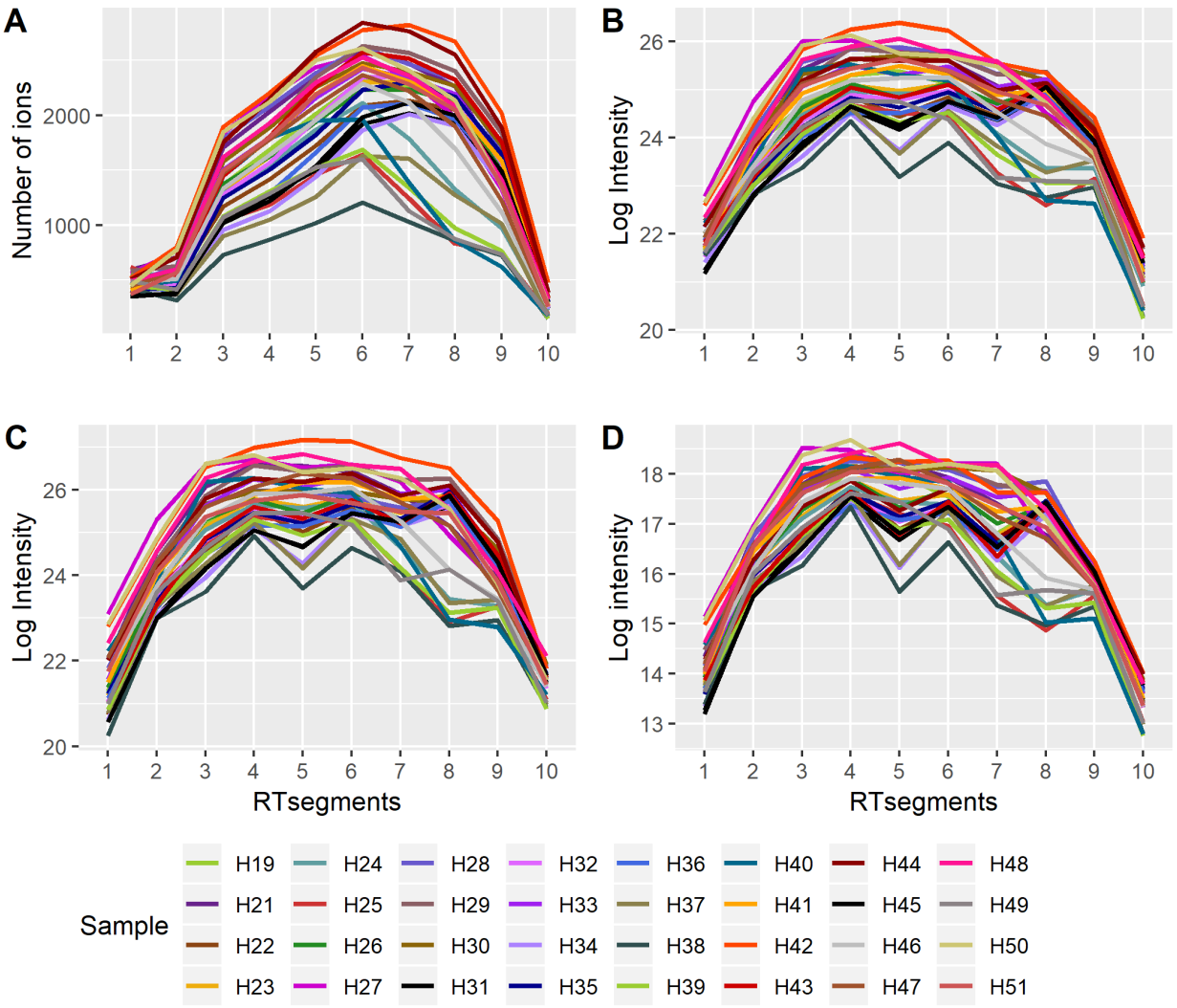


Figure 3.2 - The unpublished Steen trial dataset shows a consistent pattern amongst the samples for the A) average MS2 density over the RT segments. B) The TIC of all MS2 scans does not follow the same trend as the density with a sharper initial incline, indicating that initially, a small number of ions are making a major contribution to the intensity. C) The log of MS1TIC Total follows a very similar trend to MS2 TIC Total. D) The log transformed change in TIC from one MS2 scan to the next shows a trend similar to the TIC, indicating that ionization instability is not a problem for this instrument. The Sample number corresponds to the run order (H19-H51).

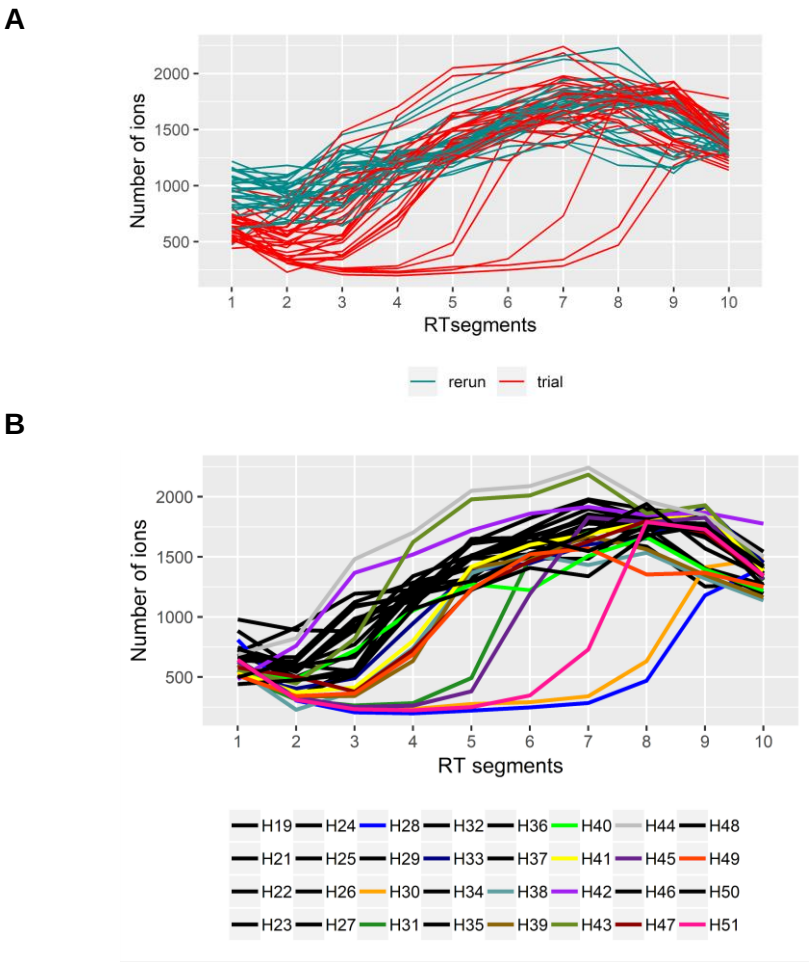
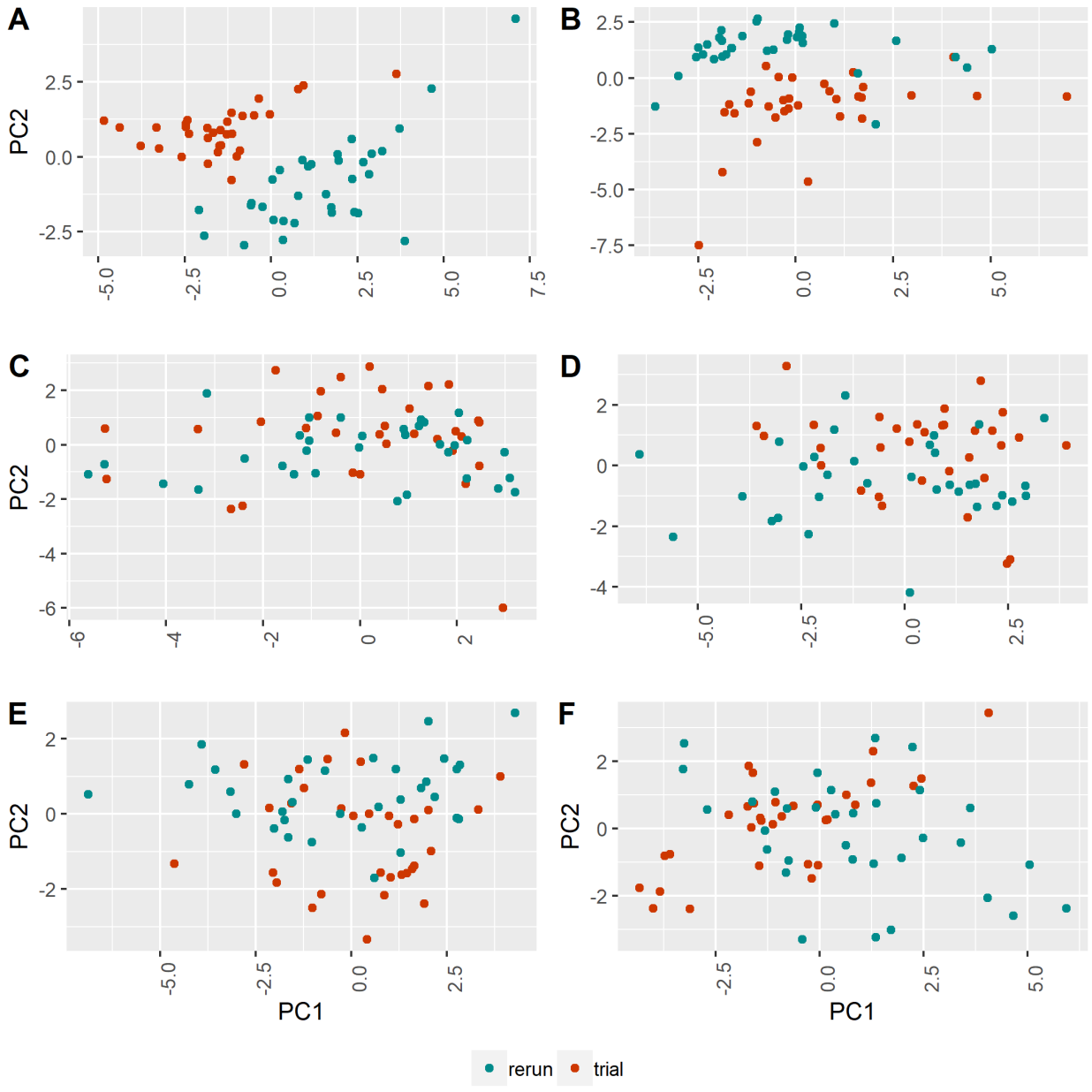


Figure 3.3A - The average MS1 ion density for the trial and rerun of the Steen dataset. Notice how the rerun dataset group into a consistent pattern, indicating that the number of ions in an ms1 scan were not supposed to show an initial drop as they did in the trial run. B - Line graph to highlight the run order of samples that deviated from the trend(coloured lines) opposed to lines that follow the trend (black). Run order was sequential for the naming structure, with H19 as the first sample and H51 as the last sample run. From this figure we note that although most out of trend samples were run in the latter half of the experiment, not all samples run close to the end of the experiment deviated from the pattern.

The visualization of a PCA of the combination of trial and rerun data was performed separately for the different RT segments (Fig. 3.4A). The first segment of the RT, the trial and rerun data separate so clearly in the space of the first two principal components that an imaginary line can be drawn between them. In the other segments the separation becomes less pronounced as the RT progresses (Fig 3.4 B-J).



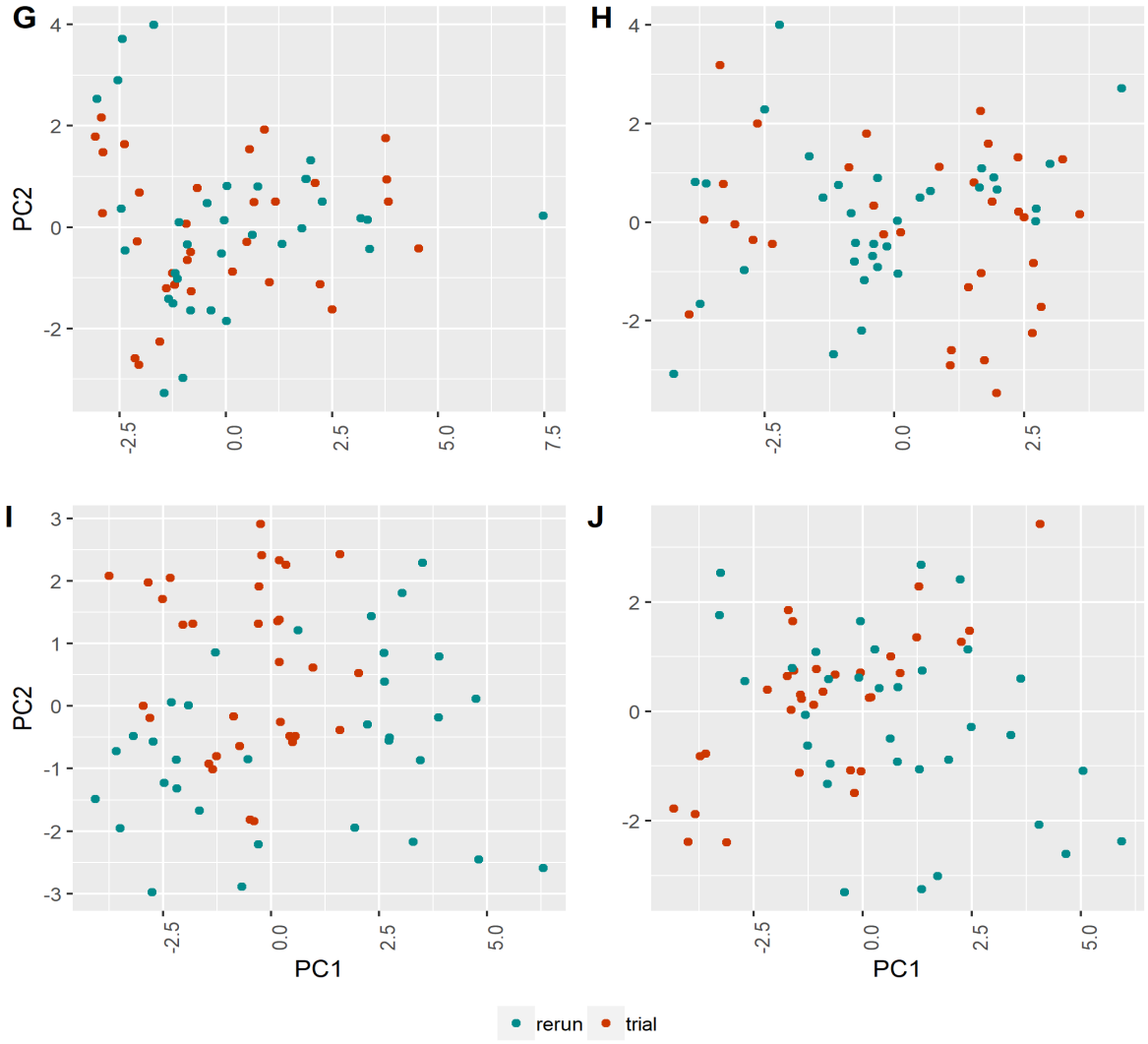


Figure 3.4 - The first two principal components of the first (A = RT segment 1) out of ten segments of the RT containing both the runs for the trial (red) and the reruns (blue). Notice the separation of the trial and rerun data. As the RT progresses (B:J = RT segments 2:10), the separation between the trial and rerun groups becomes less clear.

This data suggests that the difference between the trial and rerun samples is most pronounced at the beginning of the run and the MS1 scans of each cycle show a lower density than expected. The absence of this particular trend in the TIC metrics could be an indication that the more abundant ions in MS1 are masking the signal from less abundant ions and that

subsequently the TIC is not affected to such a large degree. The difficulty that the turbo pump was having in maintaining pressure appears to have resulted in fewer ions proceeding to the detector. The vacuum in the orbitrap compartment also changes when the HCD is switched on,[190] which could explain the problem diminishing in MS2 scans as the HCD has been switched on. SwaMe metrics were therefore able to reflect that the turbo-pump failure resulted in an inconsistency in the number of ions that were able to reach the detector, especially at the beginning of the cycle when the HCD was not yet turned on, without affecting the ionisation or TIC. This case also highlights the usefulness of segmenting the RT.

3.3.4 Scrutinizing outliers for the dataset as a whole

The Stoychev dataset was run prior to the invention of SwaMe. As part of their QC analysis of the data, several runs were identified as candidates to be rerun.

List of files rerun after manual curation of the Stoychev dataset:

003_1, 003_2, 013_1, 013_2, 056_1, 058_1, 058_2, 184_1, 184_2, 186_1, 186_2

The selection was based on comparison of MS2 fragment intensities among injection replicates (Fig.3.5) excluding anything with a CV higher than 20%, as well as the percentage of the library matched by each SWATH, with 10% as the minimum cut-off.

For this study, I decided to illustrate and compare the identification-free outlier analysis with SwaMe with their approach. For outlier analysis with SwaMe, the Stoychev dataset was rerun without dividing the RT into segments in order to include the RT metrics in the PCA for outlier analysis. PCA and distance based outlier detection revealed one outlier in the original dataset, 013_SW_2 (Fig 3.5).

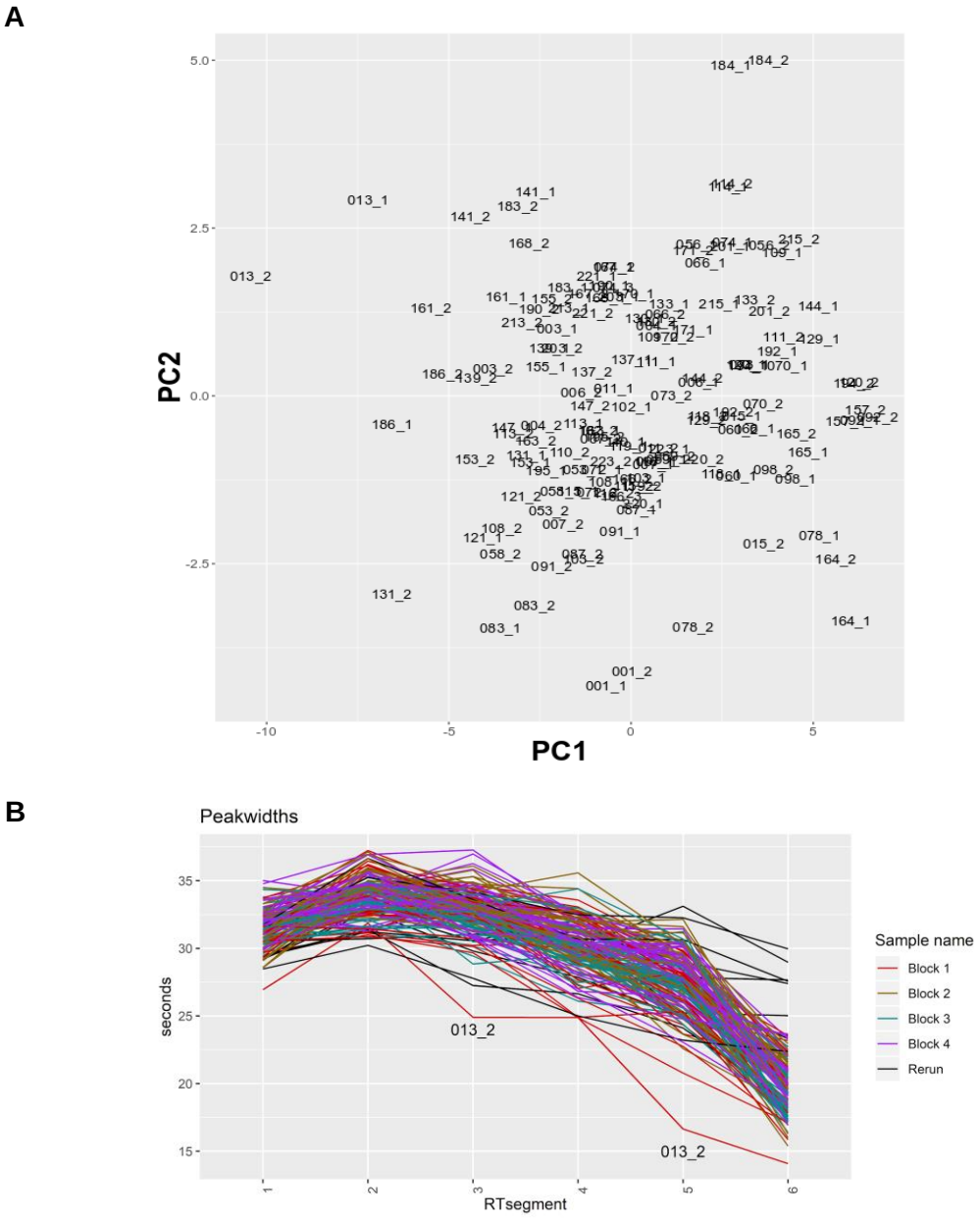


Figure 3.5A - PCA of Stoychev dataset where a sample (013_2) has been identified as an outlier. B - A line plot of the peak widths for RT segment 1-6 show this sample has a lower value for segment 2 and segment 5 than the other samples in the experiment.

Some of the metrics that proved responsible for the segregation of the outlier include chromatography metrics such as peak capacity, peak width and tailing factor, intensity related

metrics such as MS1TICTotal and MS2TICTotal, as well as density related metrics such as MS1 and MS2 average density (Fig. 3.6).

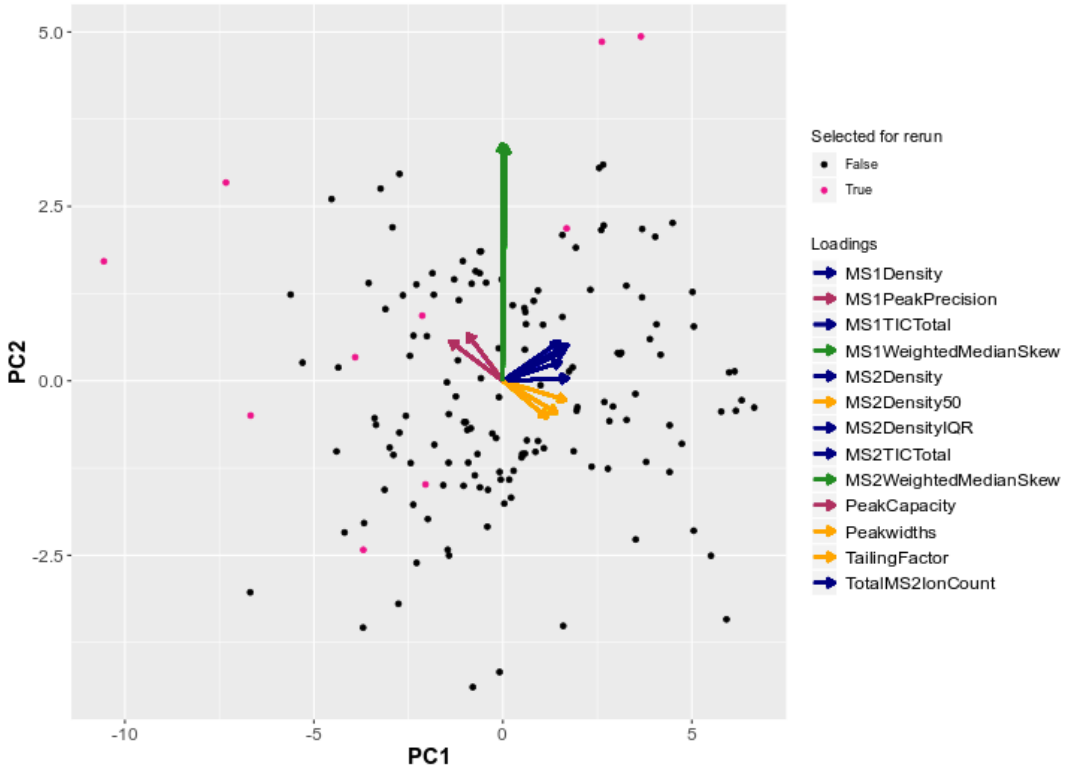


Figure 3.6 - PCA with loadings for comprehensive metrics from the Stoychev experiment.

The 11 samples identified for rerun by the Stoychev laboratory using their own approach were highlighted here in pink. The run, 013_2 that was identified as an outlier candidate in SwaMe outlier analysis was one of those selected originally by the Stoychev laboratory. Upon viewing the MS1 and MS2 total TIC (Fig. 3.7) we see that 013_2 did indeed show the lowest TIC of the samples.

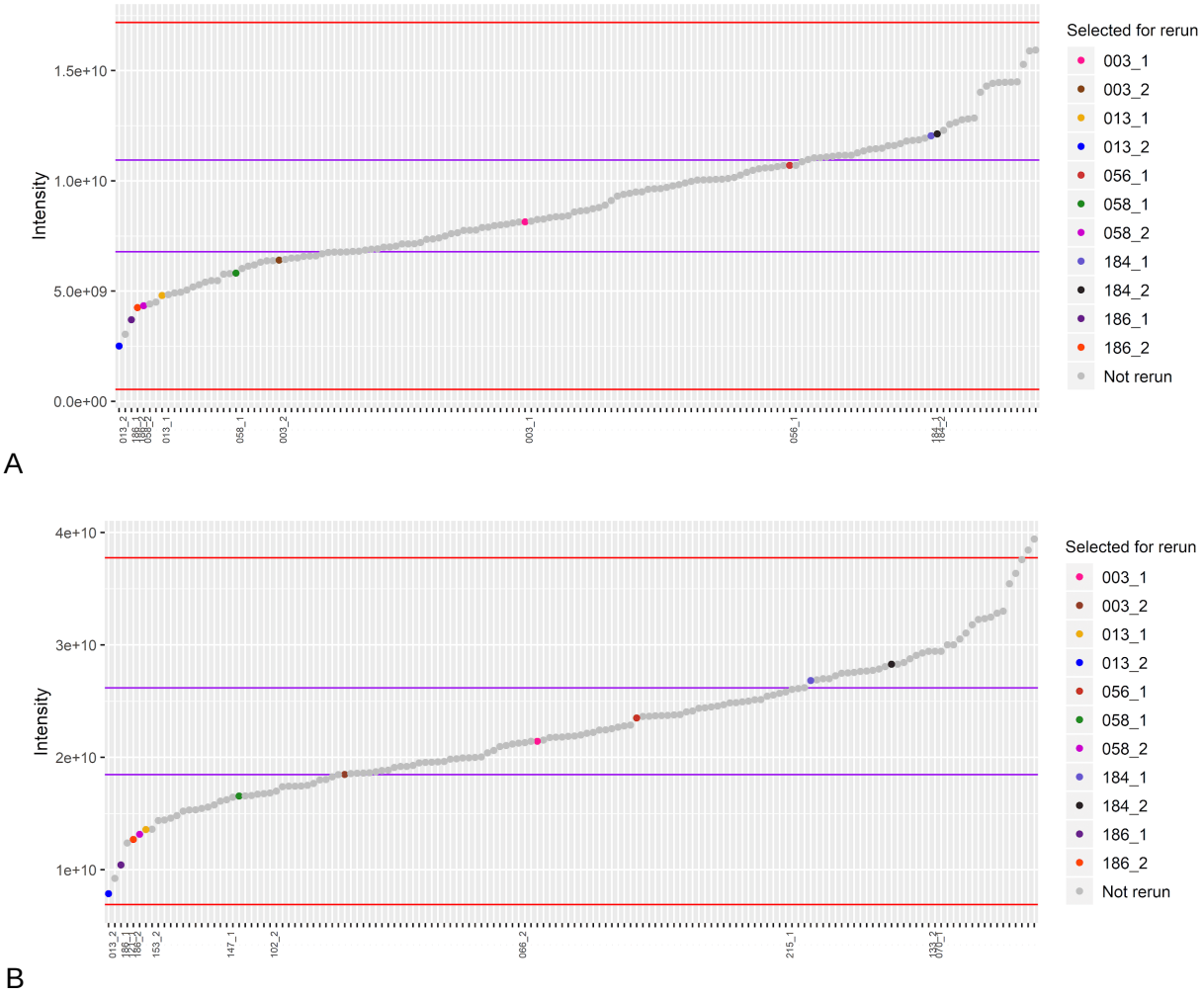


Figure 3.7 - MS1(A) and MS2TICTotal(B) from the Stoychev dataset with the samples selected for rerun displayed in blue. Purple lines indicate Q1 and Q3 and the red lines indicate Q1-1.5 x IQR and Q3+1.5 x IQR respectively. Samples 013_1, 013_2, 186_1, 186_2 showed TIC total values on the lower end of the metric, which contributed to their selection for rerun. Although 013_2 is not eligible as a possible outlier with Tukey criteria, in both cases the file has a lower value than the others.

Of the total list of samples selected for the rerun 013_1, 013_2, 186_1 and 186_2 show a lower MS1- and MS2TICTotal, as well as a lower MS2Density50, indicating that fewer ions were detected by the detector for these samples. However, when viewing other metrics such as the

average density of a MS2 scan (Fig. 3.8), the difference between 013_2 and the other samples is more pronounced and 013_2 is more than 1.5 x the interquartile range below the first quartile, thereby classifying the sample as a possible outlier.

Although MS2Peakwidths was selected as a metric contributing toward 013_2 being identified as a possible anomaly, MS2PeakWidths for the entire file did not show any potential outliers, nor did the selected samples appear to have very high values for this metric (Fig. 3.8). In SwaMe, as the software does not take identification into account, MS2 peaks are calculated for each base peak by finding occurences of the same m/z (within the mass tolerance range) within the RTTolerance (here set to 2.5 min on either side) of a base peak. Without identification data it is therefore very possible that data points could be included that do not form a part of the same individual peak.

In this case, the identification data have therefore granted the Stoychev team additional information and illuminated possible problems that did not appear in the automatic curation with SwaMe. This highlights the additional value that could be added by analyzing identification data in the quality analysis.

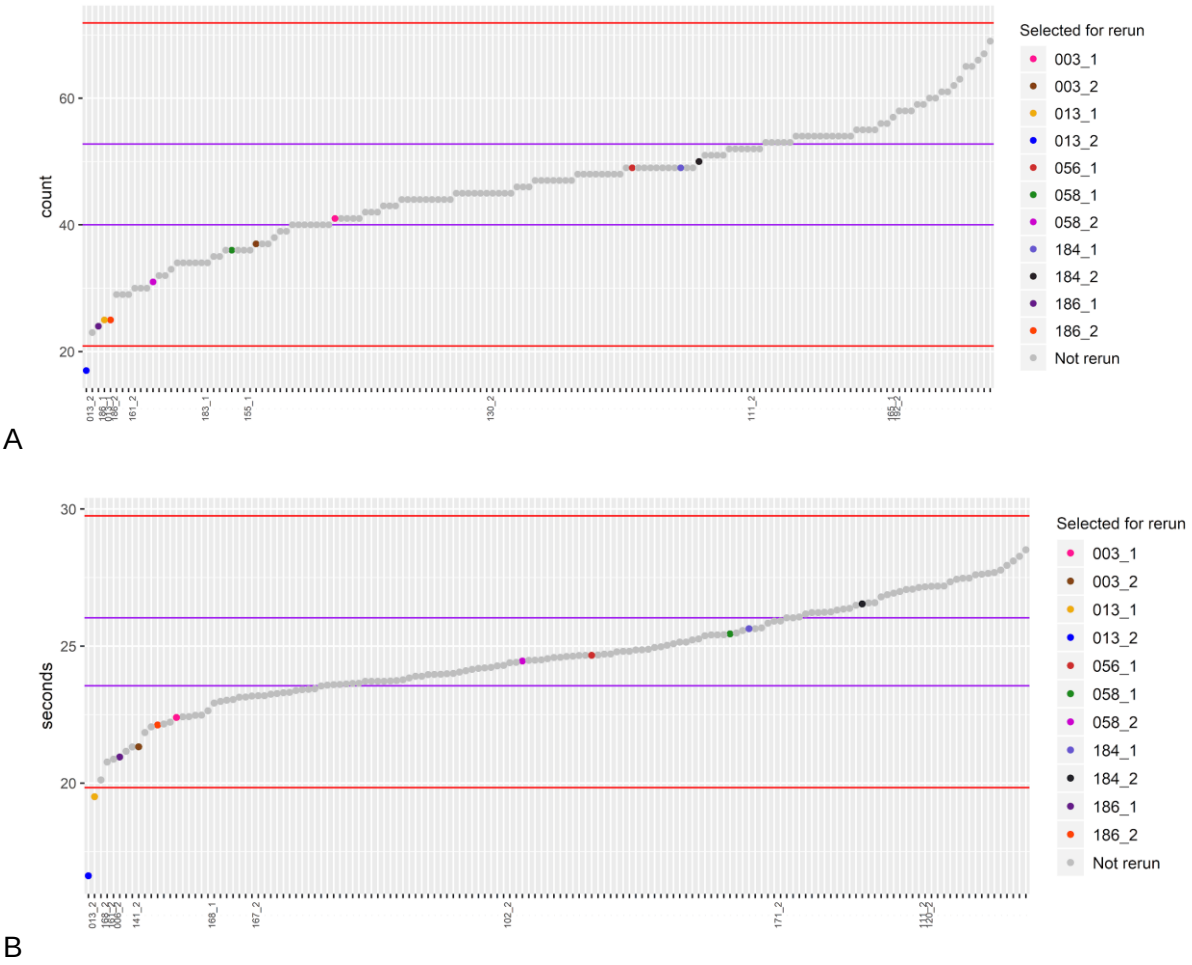


Figure 3.8 - MS2Density50(A) and MS2PeakWidths(B) and of the Stoychev dataset. Purple lines indicate Q1 and Q3, red lines indicate the Q3+1.5 x IQR and Q1-1.5 x IQR respectively to indicate possible outliers. In 013_1, 013_2, 186_1 and 186_2 the average amount of ions detected throughout the entire file was low. Note that 013_2 falls below the bottom red line in B and can therefore be considered a possible outlier in terms of its low MS2Density. The samples with low TIC and low density also showed low peak widths, which could simply be an indication that the data was sparse.

Of the metrics that are not included in the analysis due to low variance, AvgCycleTime and MissingScans should also be viewed when selecting outliers. SwaMe is the first software to

provide a metric reporting the number of empty scans in a run. Missing scans together with cycle time, peak widths and tailing factor could provide information into excessive background noise subtraction, dead volume in the system, ionisation instability, or non-specific binding of the analyte/molecule of interest to a surface of non-interest such as a plastic container or pipette tip.[192]

3.3.5 Interrogating experimental design

A well designed experiment often requires the use of blocking and randomisation.[178] Each age-race-gender matched case/control pair, in the Stoychev dataset, employed a blocking and randomisation structure in order to minimise the chance of a batch effect occurring. In a graph of the first two principal components of the RT divided quality metrics, the blocks do not separate from each other, indicating that during the experiment, the extent of instrumental drift was not excessive (Fig. 3.9). In addition, the reruns that were conducted after inspection of the peak shapes do not separate from the rest of the dataset. The lack of evidence of instrumental drift probably reflects that a little more than two months passed between the original runs and the reruns. The instrument and analyst therefore showed high reproducibility despite a batch difference.

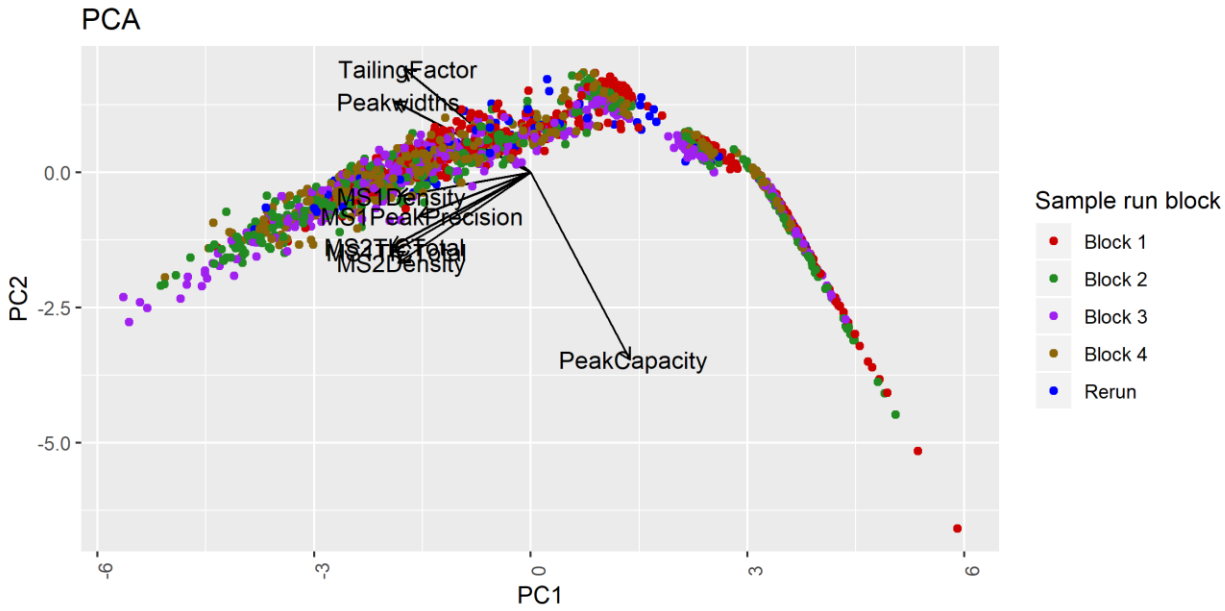


Figure 3.9 - PCA of the Stoychev dataset, showing the blocking structure. Therefore there was no evidence of a batch effect.

Furthermore, the segment of the RT that the quality metric originates from had a larger impact on the first two principal components than the origin of the sample. This is an indication of the reproducibility of the ion distribution within the RT across samples (Fig.3.10).

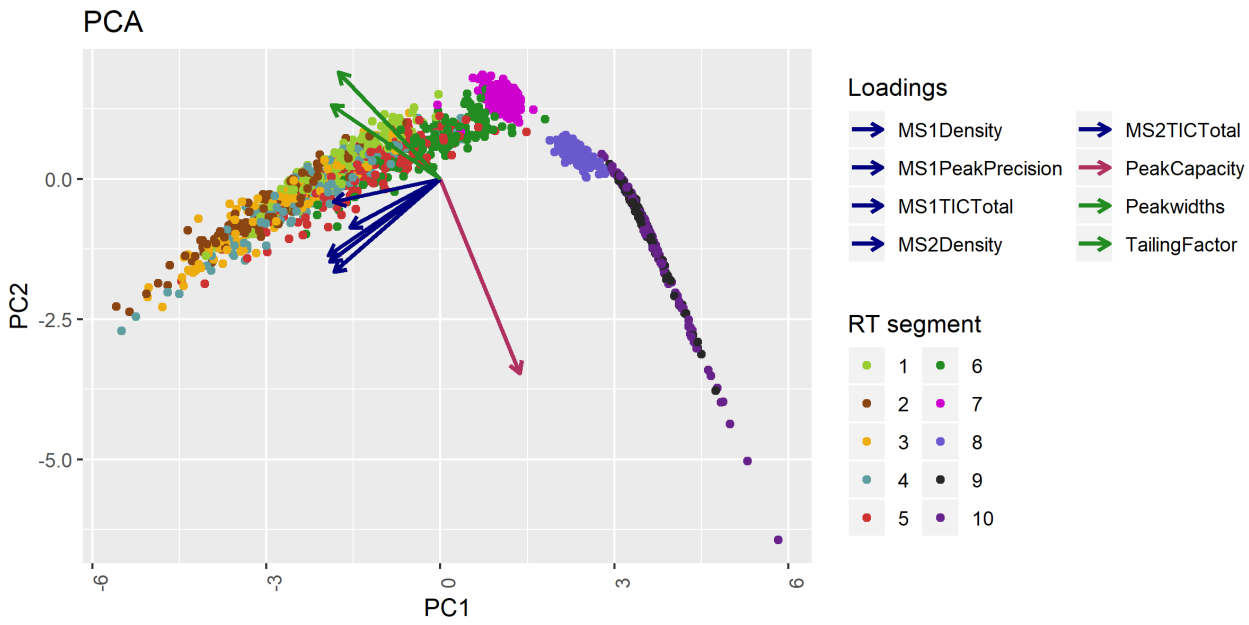


Figure 3.10 - PCA plot of Stoychev unpublished data. Samples were grouped into blocks to avoid case-control batch effects. The RT was divided into ten segments for the purposes of inspecting data quality. It is clear that the segment of the RT that the QC metrics refer to had a larger impact on the quality metrics for segments 7,8,9 and 10 than the origin of the sample as these segments separate from the first six segments in the space of the first two principal components.

In the final biological analysis of the data, the Stoychev laboratory further curated the data quality to exclude a number of samples. Here, samples were excluded if their median %CV fell above 20 or if the sample contained too much missing data.

Table 3.2 - Samples excluded from biological analysis of Stoychev dataset

Samples excluded	Reason
003	Data completeness below 10%
003 rerun	Data completeness below 10%
006	Injection replicate CV above 20%
006 rerun	Injection replicate CV above 20%
013	Injection replicate CV above 20% and data completeness below 10%
013 rerun	Data completeness below 10%
056	Injection replicate CV above 20%
056 rerun	Injection replicate CV above 20%
141	Data completeness below 10%

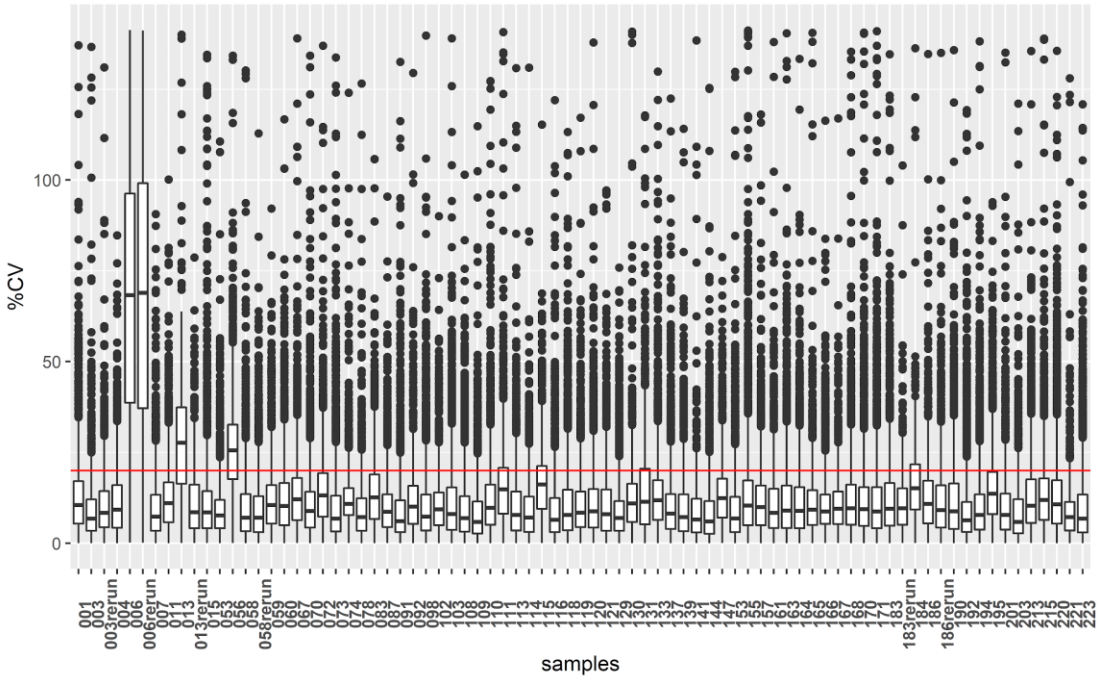


Fig 3.11 - Boxplots of the %CV of MS2 fragment intensities between injection replicates, produced from CV's computed from Spectronaut 14.[191]

3.3.6 Abundant ions masking signal

Upon comparing the MS2TICTotal and the MS2 ion density trends of the Aebersold dataset,[182] we are able to view similarities for the number of ions and the intensity of signal generated (Fig. 3.12). This similarity becomes important in identifying segments of the RT where a small number of ions with high intensity may be masking the signal of other ions and depletion or a change in gradient may be considered. In such sections, it is also possible that excessive background noise subtraction may have led to a reduction in sensitivity.

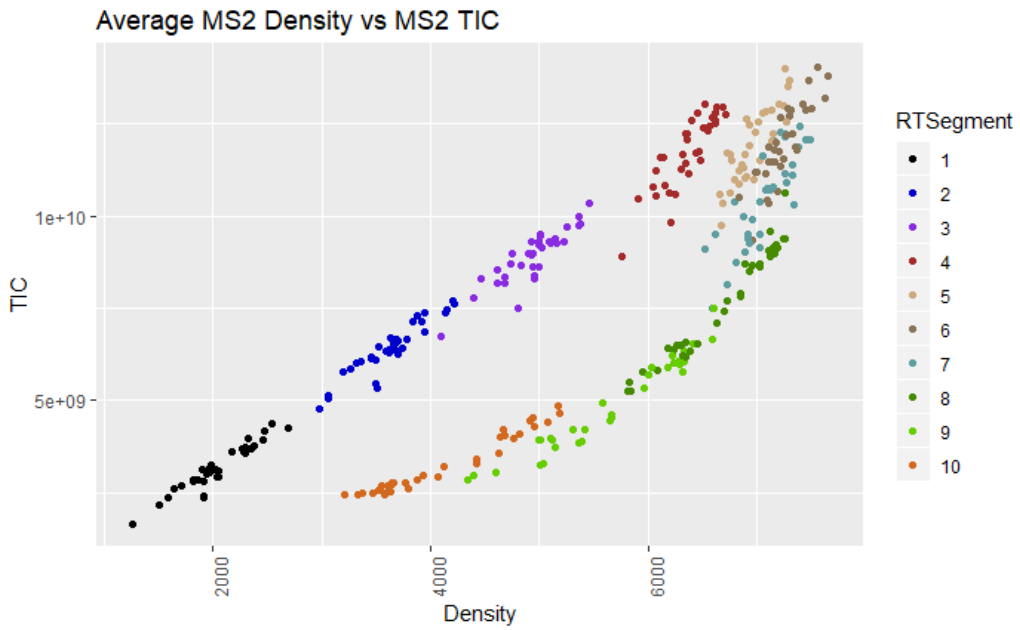


Figure 3.12 - The average MS2 density plotted against the total MS2 TIC per RT segment of the Aebersold dataset. If the majority of the TIC can be explained by a few ions only, the segment is at high risk of a few abundant ions masking signals from less abundant ions. Therefore, for the first few segments of the RT, fewer ions contributed to the TIC than in the last few segments, however the trend is still following a linear projection indicating the TIC to density ratio is rising.

3.3.7 Analysing Waters MS^E data

The SCIEX TripleTOF and Thermo Fisher Q-Exactive instruments are very specific in the number of scans that are collected and the instrument does not deviate from the settings input by the analyst. This is not the case in Waters data, where every MS1 scan is not always accompanied by a tandem mass spectrum and the tandem mass spectra involves only a single window for the entire dynamic range. As a result, we handled the data from this set similar to the way we would for the other instruments if only one very wide SWATH was present. In the Pereira dataset,[12] the swath divided metrics in most cases become a repetition of the comprehensive metrics as each cycle contains one window at most and the swath-divided metrics therefore only provide one value. For example, MS2Density50 from the comprehensive set of metrics provides the average density for all MS2 scans. The metric, swDensityAvg from the swath divided metrics returns for each unique isolation window target, the average ion density. As all MS2 scans now have the same isolation window target, the two values are equal. In this dataset, the quality metrics were influenced more by the time that elapsed after incubation than their run order or their case/control status (Fig. 3.13). Where experimental drift may have influenced the runs in temporal proximity, the nature of the proteins created by the fungus at different growth stages may have had such a large effect in this case that instrumental changes are overshadowed. This dataset is an example of biological findings showing in quality control data.

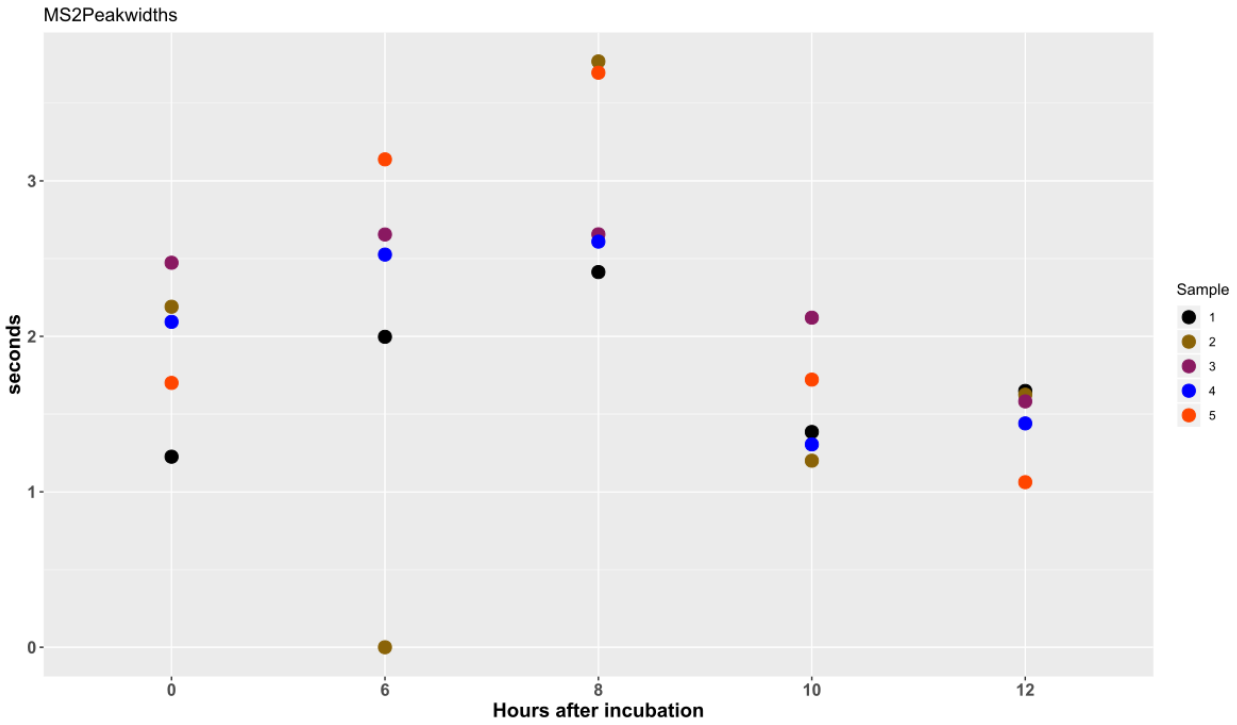


Figure 3.13 - Both cases and controls were run in sequential fashion, so that all the timepoints of one sample were completed before another began. The MS2PeakWidths for samples of the fungus, Paracoccidioides lutzii, was more similar between different growth stages for the colony than they were between runs of similar temporal proximity for the Pereira set[184] which was collected using Waters MS^E.

3.4 Conclusion

In this study, experimental and QC DIA datasets from instruments from three different vendors were analyzed with the quality control metric generating software, SwaMe. The analysis of experimental datasets with PCA illustrates the impact of isolation schemes on quality metrics, as well as the importance of block design. In addition, post-hoc analysis of a dataset halted due to a turbo pump failure shows the increased variability detectable in quality metrics.

Chapter 4: Assurance - downstream analysis of biological mass spectrometry quality metrics

4.1 Introduction

Command line MS/MS quality tools QuaMeter,[99] QuaMeter ID-Free,[171] and a new DIA version, SwaMe, discussed in chapter three provide valuable quality metrics for discovery proteomics. These metrics can support decision making,[99,168,171] but this requires the implementation of a statistical model in a language such as R or Python. A plethora of free tools available for proteomic analysis have allowed bench biochemists to interpret experiments without using statistical languages or tools.[103,193–195] As a result, one cannot assume statistical programming skills in even Ph.D.-trained proteomics researchers. Despite the existence of multiple other MS quality tools for bench biochemists that are unrestricted to programming skills,[106,108,112] none can analyze metrics produced by QuaMeter or SwaMe. In addition, the command line nature of QuaMeter can prove daunting to many biologists.

Unfortunately, these hindrances may turn bench biologists away from in-depth quality analysis. There is therefore a need to be able to both run command line quality software such as QuaMeter and SwaMe through a user interface as well as perform downstream statistical analysis similar to that of chapters 2 and 3. The increased accessibility that such a tool provides enables a complete pipeline for proteomic analysis and will promote the use of quality control software in the field of proteomics.

The aim of this chapter is to create a tool for the downstream analysis of QuaMeter and SwaMe metrics which is able to perform unsupervised outlier detection, supervised classification and

display the distribution within each metric as well as run the command line tool through its user interface.

4.2 Materials and methods

4.2.1 Datasets

Table 4.1 - Datasets included in the study

Last author on dataset paper	**Accession number**	**Instrument**	**Description**
Smith[168]	PXD000320, PXD000321, PXD000322, PXD000323, PXD000324	Thermo Fisher LTQ Orbi-XL	DDA: 76 manually curated files used for test and training, 21 manually curated files used for analysis, mzIdentML files are also available for all runs
Tabb[142]	PXD006843	Orbitrap Velos	DDA: 120 raw files(previously shown in chapter 2 to have 2-4 quality outliers depending on the method)

The datasets used in the analysis include one clinical dataset and one longitudinal dataset.

For random forest analysis, data from the instrument, Eagle, was extracted from the Smith dataset and randomly divided into two sets - the analysis set and a set that would later be divided into the test and training set by Assurance. The .mzid files of the test and training set created by Amidan and colleagues (2014) were also used to allow classification of bad quality data via identification results and from a table of the quality metrics. The test and training set

quality data has been manually curated by experts in the field.[168] This curation was used to annotate “good” and “bad” data by uploading a .tsv file of the QuaMeter ID-Free metrics (Available at https://github.com/marinaPauw/Assurance/releases). This approach was compared to classifying “bad” data on the basis of their identification data from mzidentMLs created as a part of the original analysis[168] via MSGF+. [166]

Assurance offers two separate methods for a manual classification of the samples into “good” and “bad” quality. The first involves uploading identification files and selecting “bad” quality files from a graph of the spectral counts. The other involves the uploading of a tsv file of the quality metrics and selecting rows from this table as “bad” quality.

4.2.2 Explanation of Assurance structure

The main window provides the option of either running QuaMeter, SwaMe or uploading previously produced results (Fig 4.1). Arguments for running the tools can be adjusted to reflect instrument type, and the terminal output is displayed in a window. If SwaMe or QuaMeter is run, output files are automatically loaded upon completion. The quality metrics can then be used to detect outliers and/or interrogate each metric separately and/or perform random forest analysis for longitudinal data. This last requires the upload of additional data that will be balanced and randomly divided into a test and training set for classifier evaluation as well as allocating the good or bad quality data within the group.

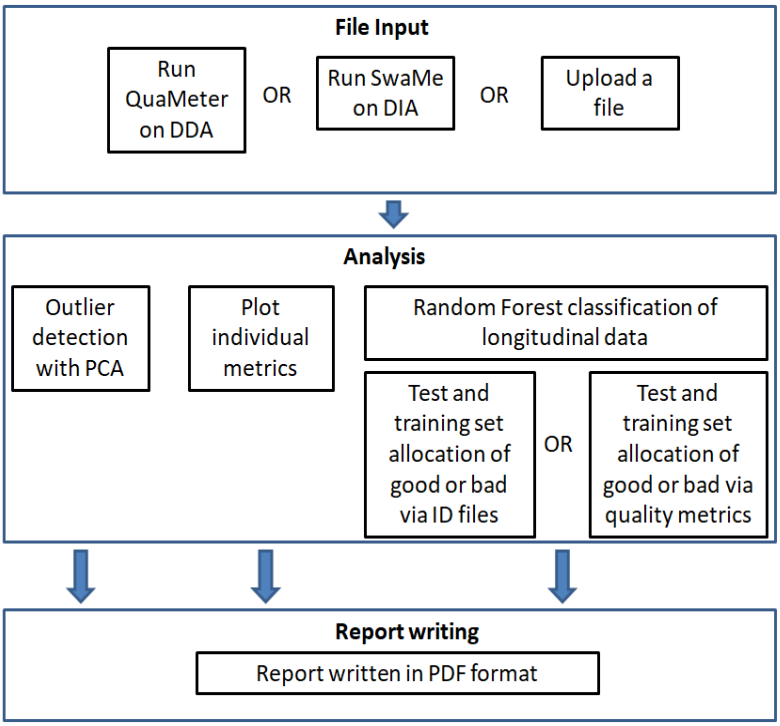


Figure 4.1 - Flow diagram representing an Assurance run.

4.2.2.1 Unsupervised outlier detection via PCA

For the PCA Assurance employs the Scikit-learn[197] package. Non-numeric columns and columns with less than 1% variance are removed. In the case of columns that are more than 99% correlated, the tool selects the first column excludes the others. The data is scaled and PCA performed. Thereafter, the tool determines the number of significant components using the Elbow method and a distance matrix is created. The median of the distances for each run is calculated and possible (1.5 x IQR above Q3) and probable outliers (3 x IQR above Q3) of the medians are indicated in blue and red respectively.[196] For easier reading, in the rest of the chapter, possible outliers have been abbreviated to “PossOut” and probable outliers to “ProbOut”. The data excluding ProbOuts can be reanalysed to determine if initially discovered outliers are masking others. The metric loadings can also be displayed.

4.2.2.2 Visualisation of metrics

Individual metrics are arranged in order of ascending value and plotted as a line graph. In addition, for numerical metrics the first(Q1) and third(Q3) quartile as well as Tukey's designation for possible outliers (Q1+1.5 x the interquartile range(IQR) and Q1-1.5 x IQR)[196] is represented as horizontal lines.

4.2.2.3 Longitudinal analysis via Random Forest

The execution of random forest analysis involves the h2o package.[198] The package runs the h2o jar executable via Java and creates a local server enabling the rapid analysis of large quantities of data. Additional data in the form of a training and test set must be uploaded. Classification of this set can occur via the uploading of identification files and selecting from a subsequent graph. The allowed file formats are .mzid[199], .pepXML and MaxQuant tab-delimited summary statistics.[193] Alternatively, a table of the quality metrics can be used for classification. The dataset is randomly divided into training and test data and hyper parameterization used to create a model. The model is then trained, tested and used for classifying the analysis data. Performance metrics of the test set, metric contribution and the probability of being classified as 'bad' for each sample is displayed.

4.2.2.3.1 Classifying the training set from quality metrics

The "Curated Quality" column (Fig. 4.2) in the test and training set data was used to select all the samples. The samples for this set had been manually curated by experts in the field based on identifications, peak shape and other characteristics into three categories: "Poor", "ok", and "good". This column is present in the quality data downloaded as part of the supported materials.[168]

For the purposes of our work, we will consider both "ok" and "good" as samples with good quality and we will consider "poor" as samples with bad quality. In the set used for test and

training, there were 33 samples of “ok” and “good” quality and 43 samples of “poor” quality. If a researcher has prior knowledge on the quality of data for the test and training set as is expected with longitudinal data, a column can be safely added to the QuaMeter .tsv file to discern this difference within Assurance and to increase the process repeatability. As long as the column heading is not also present in the analysis data, it will not be included in training the model.

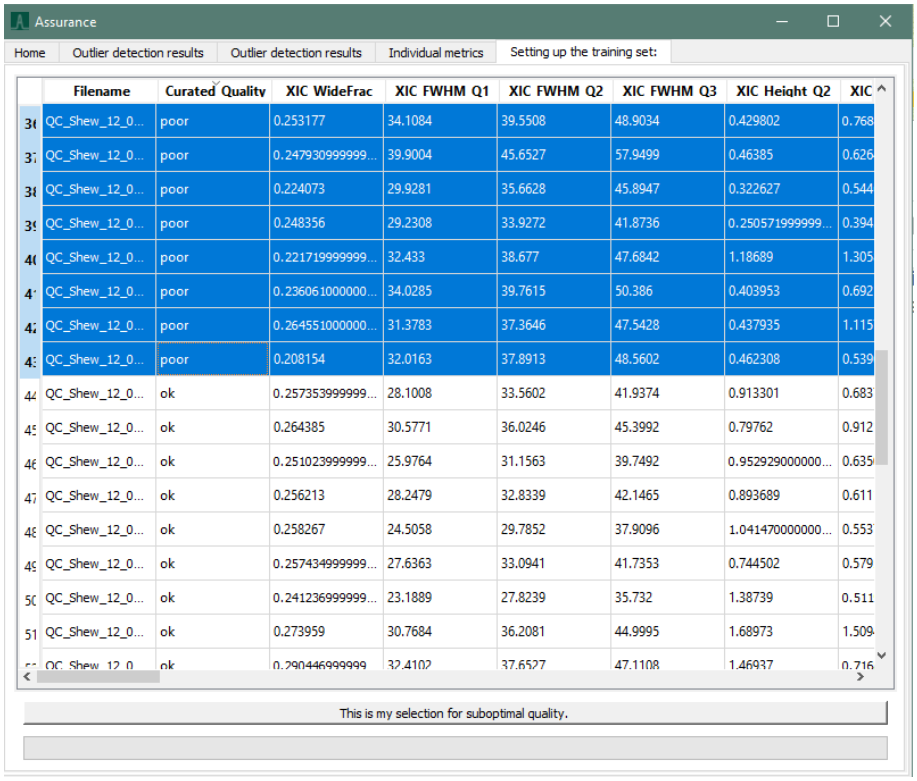


Figure 4.2: Screenshot of the selection of the poor quality data from the EagleTrainingandTest.tsv available at https://github.com/marinaPauw/Assurance/releases.

4.2.2.3.2 Classifying the training and test set from the number of spectral IDs

As an alternative strategy, the identification results were uploaded and the 28 samples with the lowest spectral counts were annotated as ‘bad’ due to a natural grouping in the dataset (Fig 4.3). However, in the case of a researcher that does not have any indication of which data points constitute poor quality, ‘bad’ quality data could be allocated on a more quantitative basis

such as data in the bottom quartile of the identification distribution or data that make up the average -1.5x IQR. In the case of the Tabb dataset. After the allocation, the quality files were uploaded. The quality files are then used to train and test the model and the identification files are excluded from the rest of the analysis.

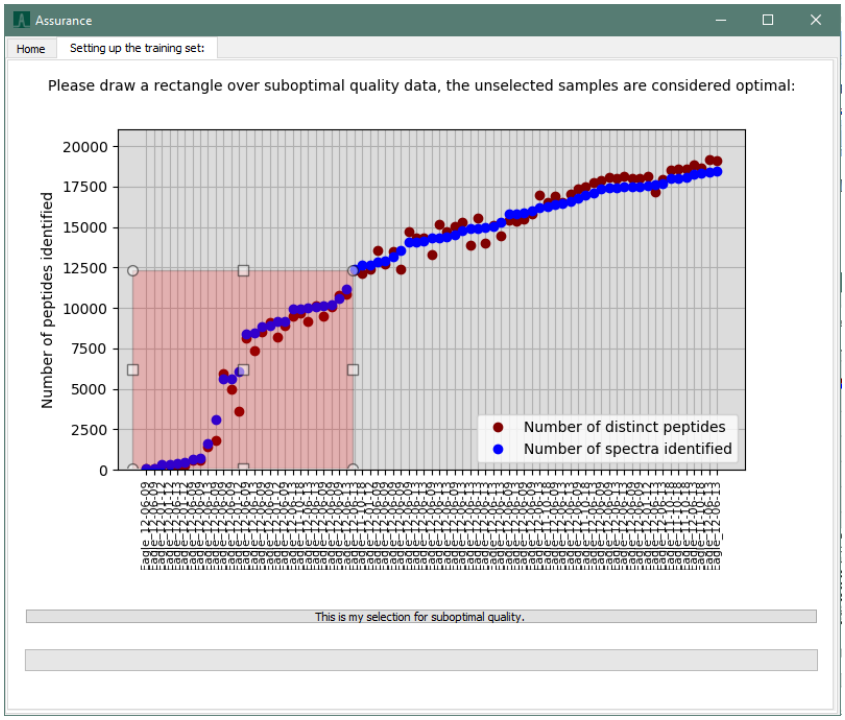


Figure 4.3: Screenshot of the selection of 'bad' quality files from the Smith dataset as the 28 samples with the lowest number of identified spectra.

Whichever method is chosen, it is therefore of utmost importance that the test and training set be correctly classified. It is advisable to run the test and training set through the outlier detection to ascertain that there were no quality outliers in the "good" quality section.

4.2.3 Report generation

After running one or more functions, the software can generate a pdf report. Accreditation in the medical and food industries require the quality control team to keep record of the instrument

quality in the form of reports. Often this entails storing both hard copy and electronic versions for audits and for review in the case of a customer complaint or a product quality query. The PDF creation function of Assurance is easily printed and stored electronically to fulfill quality accreditation requirements.

4.3 Results and Discussion

4.3.1 Outlier analysis on the Tabb dataset

Upon performing PCA on the Tabb dataset (Fig 4.4), two data points that had previously been identified as anomalies in chapter 2 are again noted (SW2-1-9, SW2-1-10). However, six additional samples are noted as ProbOuts (>= 1.5 x IQR from Q1/Q3) in red and five more are noted as PossOuts(>= 3 x IQR from Q1/Q3) in blue (Fig 4.4).

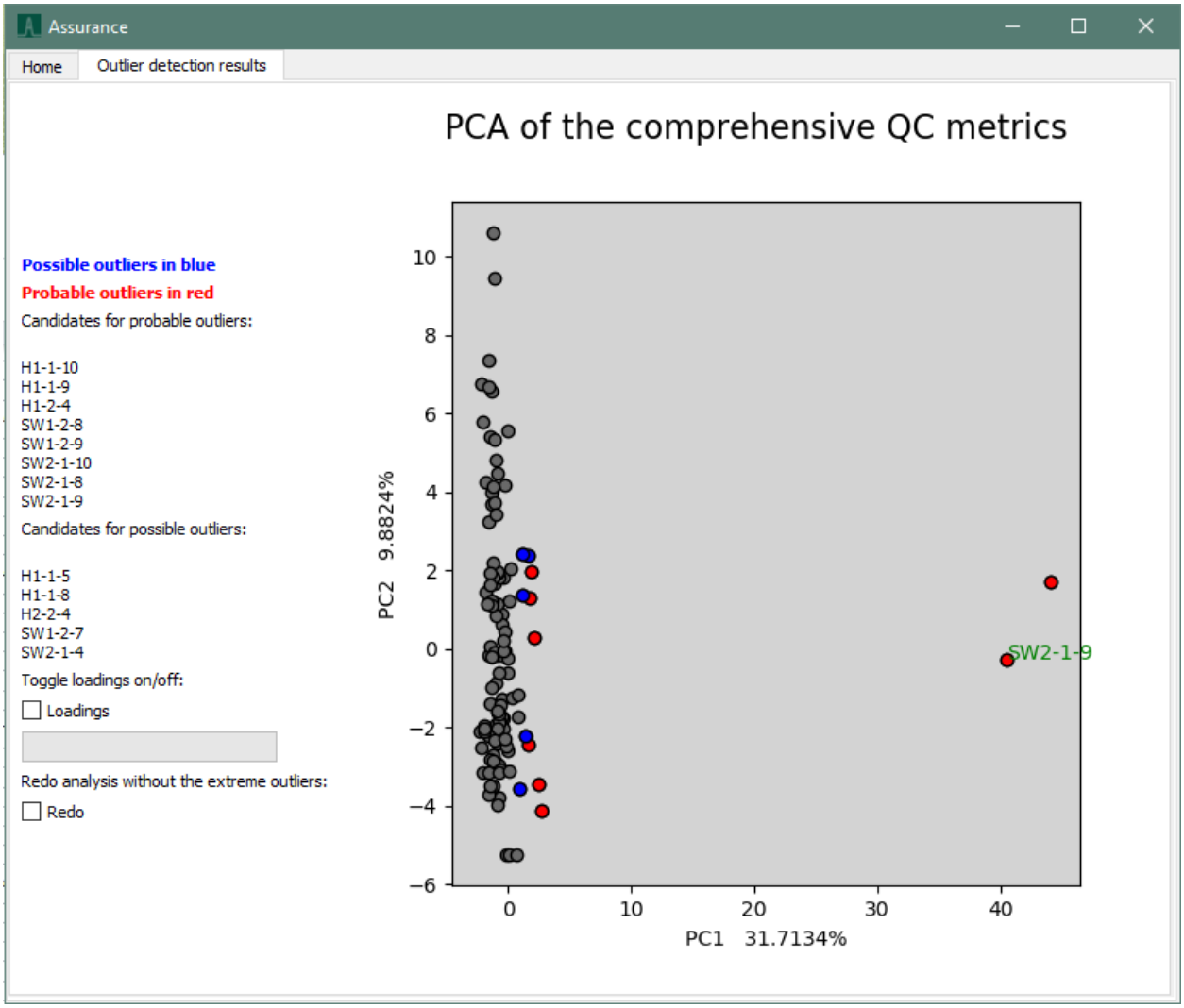


Figure 4.4: Screenshot of the outlier analysis results for the Tabb metrics. The ProbOuts are marked as red data points and the PossOuts are marked in blue. To the left of the PCA plot, the PossOuts and ProbOuts have also been listed.

This discrepancy between the results noted here and those of chapter two (Fig. 2.2) is the consequence of a slight difference in methods. During analysis in chapter two, a robust PCA was constructed in R. Given that Assurance may be employed on sets containing fewer experiments, the software was developed around a conventional PCA computed with sklearn from Python instead. It is therefore apparent in the dataset in question that the aberrations have

had a large effect on the PCA. This is also visible when noting the other PossOuts and ProbOuts are the closest to the outlier samples on the first axis, indicating that the weight of the two most obvious anomalies had a large impact on the positioning of data points in the first two principal components.

By toggling the loadings checkbox, it becomes apparent which metrics are responsible for the sample positioning in the graph (Fig. 4.5).

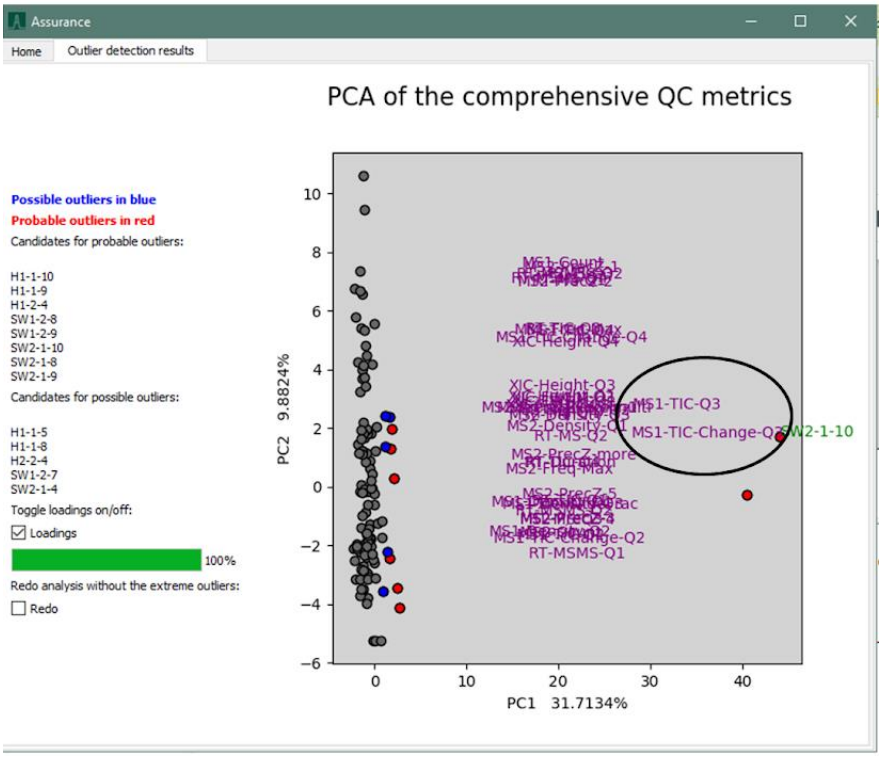


Figure 4.5: Screenshot of the outlier analysis results for the Tabb metrics with the loadings annotated in purple. MS1-TIC-Q3 and MS1-TIC-Change-Q3 notably increase in the direction of the two most prominent anomalies and for clarity they have been circled.

Due to the large distance between the two prominent aberrations and the bulk of the data, it is very possible that additional outliers in the dataset may be masked by the effect that the anomalies have on the data positioning in the realm of the significant principal components.

After re-analysis without the ProbOuts, two new ProbOuts are noted. The six PossOuts also differ from the first analysis.

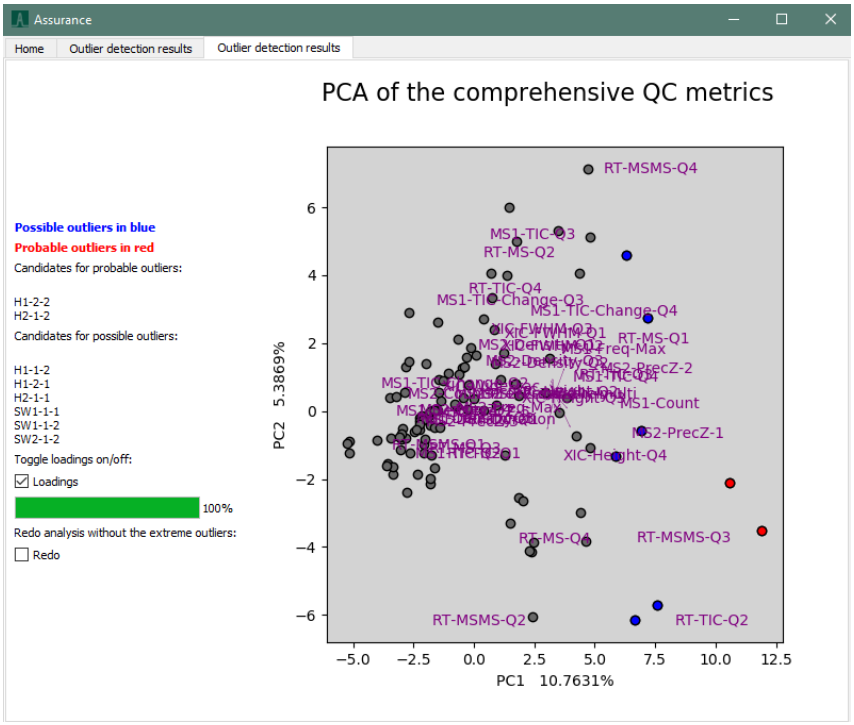


Figure 4.6: Screenshot of the anomaly analysis results for the Tabb metrics after reanalysis with the metric contributions annotated. For this reanalysis, the ProbOuts and PossOuts were not noted in the previous analysis and metrics RT-TIC-Q2, RT-MSMS-Q3 and MS2-PrecZ-1 increase in the direction of the anomalies, with MS1-TIC-Q3 and MS1-TIC-Change-Q3 decreasing.

4.3.2 Individual metrics of the Tabb dataset

The very first metric in the individual metrics section, StartTimeStamp, illustrates the 23-day gap in the data noted in chapter 2 (Fig.4.7). This project demonstrates the value of including StartTimeStamp in the quality analysis.

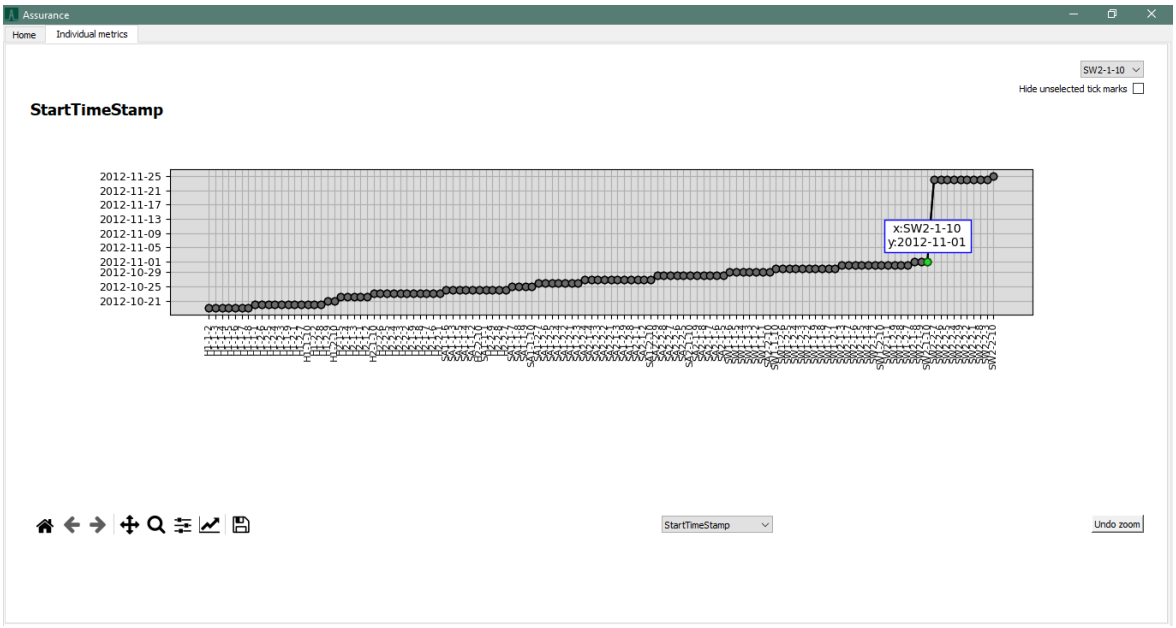


Figure 4.7: Screenshot of the first individual metric, runDate. Samples were run sequentially until a marked 23-day gap occurred. Noticeably, the gap occurred directly after the two most prominent outliers were analysed. The selected sample - indicated in green, can be changed by clicking a data point or via the dropdown menu at the top-right.

When viewing the metrics indicated to contribute to the status of the ProbOuts, such as MS1-TIC-Q3, it becomes clear that the two most prominent anomalies were indeed a large distance from the rest of the sample distribution. In addition, the other samples identified as ProbOuts before the reanalysis are visible here as the samples with the next six highest values. This is again indicative that the metrics in which the two most prominent aberrations were outliers played a very large role on the principal components and may indicate that the additional ProbOuts in this dataset may not be true anomalies. When inspecting the distribution of this dataset it is clear that sample H-1-2-4 has a value greater than the Q3+1.5x IQR as is indicated by the blue line.

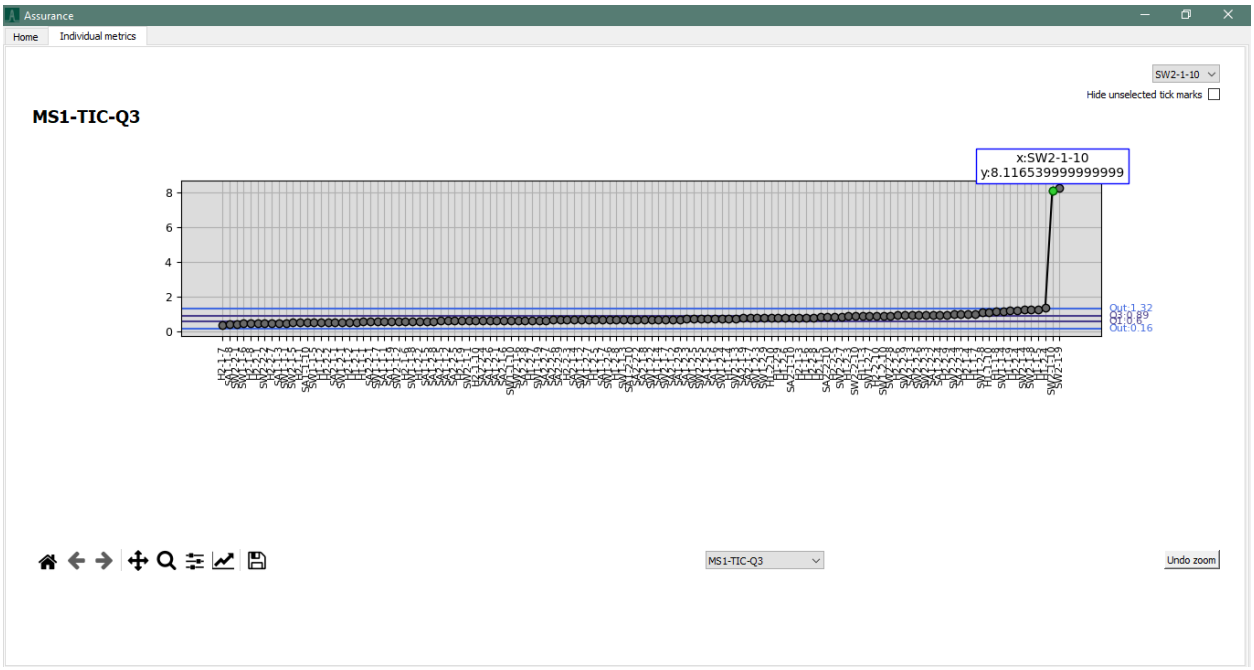


Figure 4.8: Screenshot of MS1-TIC-Q3 for the Tabb dataset. Note the two most prominent ProbOuts are located far above the blue line depicting Q3 + 1.5 x IQR, marked in the graph as "Out: 1.32". The other samples that had been classified as ProbOuts follow as the next highest values for this metric.

4.3.3 Random Forest analysis of Smith dataset

This dataset was manually curated by experts as "good", "okay" or "poor" based on identification results. The test and training set for the random forest analysis was classified in one of two ways. In the first case, the quality metrics were combined into a table and a column was added with the manual curation results of test and training set. The data curated as "poor" was allocated to the "bad" quality group and the data labelled "okay" and "good" during curation were designated "good" data. The second strategy involved viewing the identification results and selecting poor quality based on the resulting graph.

4.3.3.1 Comparison of the two classification strategies

The model resulting from the quality table classification strategy showed an accuracy of 54.84% when tested on the testing data and classified 16 of the 21 samples as 'bad'(Table 4.2). The proportion of trees that voted each sample as bad is noted in Fig 4.9.

Table 4. 2: Comparison of the two methods with the manual three level classification system of experts in the original article [3] Manual curation of "ok" and "good" quality were both considered to be "good" in our binary classification system, where "poor" was considered to be the same as "bad". Where the model's predictions agreed with that of the manual curation the cell was coloured green(true positive (TP) and true negative (TN)), else red (false positive (FP) and false negative).

SampleName	**Manual curation from Amidan and colleagues[168]**	**ID predicts**	**Quality predicts**
QC_Shew_12_02_Run-14_6Sep12_Eagle_12-06-13	good	TP	FN
QC_Shew_12_02_Run-16_6Sep12_Eagle_11-10-18	good	TP	TP
QC_Shew_12_02_Run-13_28Aug12_Eagle_12-06-09	ok	TP	FN
QC_Shew_12_02_Run-08_6Sep12_Eagle_11-10-18	ok	FN	TP

QC_Shew_12_02_Run-09_28Aug12_Eagle_12-06-09	ok	TP	FN
QC_Shew_12_02_Run-12_5Sep12_Eagle_12-06-13	ok	TP	FN
QC_Shew_12_02_Run-02_28Aug12_Eagle_12-06-13	ok	TP	FN
QC_Shew_12_02_Run-10_22Aug12_Eagle_12-06-13	poor	TN	TN
QC_Shew_12_02_Run-06_23Aug12_Eagle_12-06-13	poor	FP	TN
QC_Shew_12_02_Run-05_28Aug12_Eagle_12-06-09	poor	TN	TN
QC_Shew_12_02_Col-4_Run-12_18Aug12_Eagle_11-10-18	poor	TN	FP
QC_Shew_12_02_Run-04a_25Sep12_Eagle_11-10-18	poor	FP	TN
QC_Shew_12_02_Run-7_22Aug12_Eagle_12-06-09	poor	TN	TN
QC_Shew_12_02_Run-05_24Sep12_Eagle_12-06-09	poor	FP	TN
QC_Shew_12_02_Col-4_Run-4_18Aug12_Eagle_11-10-18	poor	TN	TN
QC_Shew_12_02_Run-10a_25Sep12_Eagle_12-06-13	poor	FP	TN
QC_Shew_12_02_Run-9_22Aug12_Eagle_12-06-09	poor	TN	TN
QC_Shew_12_02_Run-14_23Aug12_Eagle_12-06-13	poor	TN	FP
QC_Shew_12_02_Run-01_30Aug12_Eagle_12-06-09	poor	TN	TN
QC_Shew_12_02_Run-15_5Sep12_Eagle_12-06-09	poor	FP	TN
QC_Shew_12_02_Run-05_23Aug12_Eagle_12-06-09	poor	FP	FP

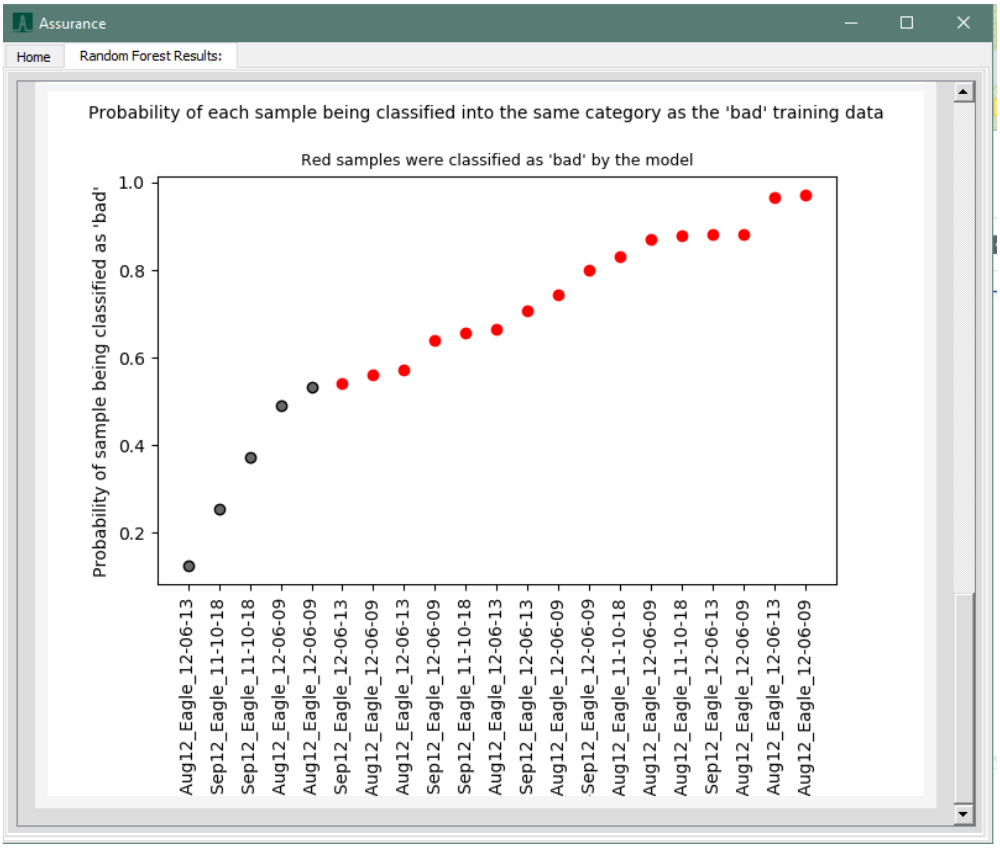


Figure 4.9: Screenshot of the proportion of trees that voted each sample in the quality metrics file classification round as 'bad'. The samples classified as bad are marked in red.

The metric contribution indicated RT-TIC-Q1 and MS1-TIC-Q2 as the two main contributing factors (Fig 4.10).

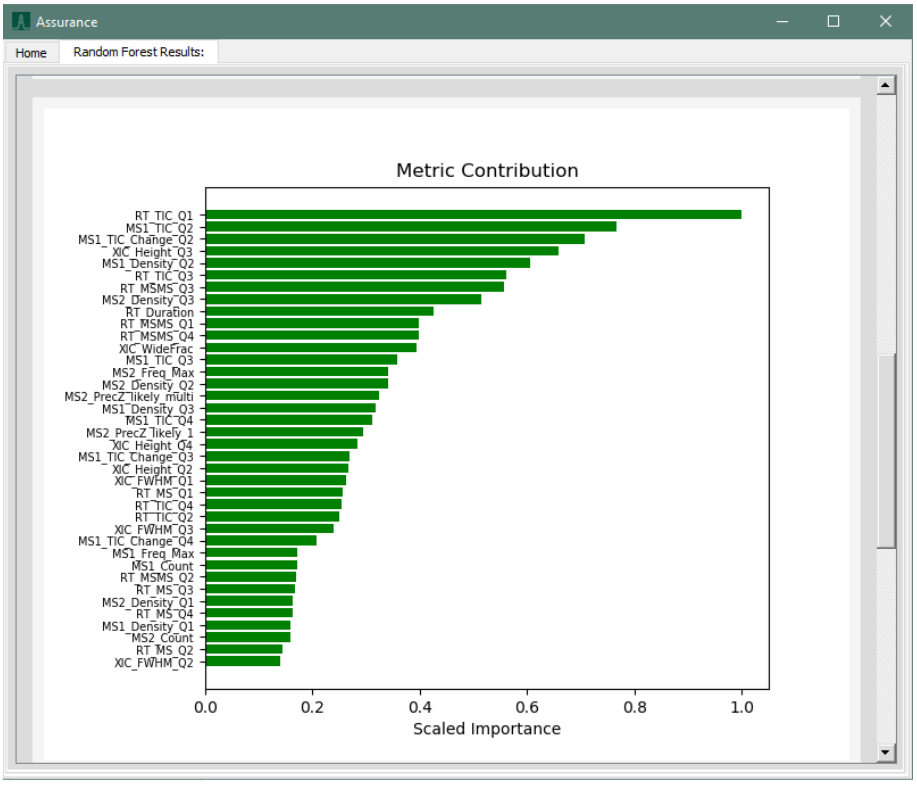


Figure 4.10: Screenshot of the metric contribution for the random forest analysis via table of the quality metrics.

On the other hand, the model that was classified from the graph of the identification files showed 76.67% accuracy on the test set and 9 samples were identified as ‘bad’ quality. The plot representing the proportion of trees that voted each sample as bad is displayed in Fig. 4.11. The metric contribution is displayed in Fig 4.12. The metrics contribution order between the two techniques is clearly very different, with RT-TIC-Q3, the highest contributing metric when the quality table was used, in 31st place when the identification graph is used. RT-MSMS-Q2 which was 32nd when the quality metrics were used is first when the graph of identification metrics is used (Fig 4.12).

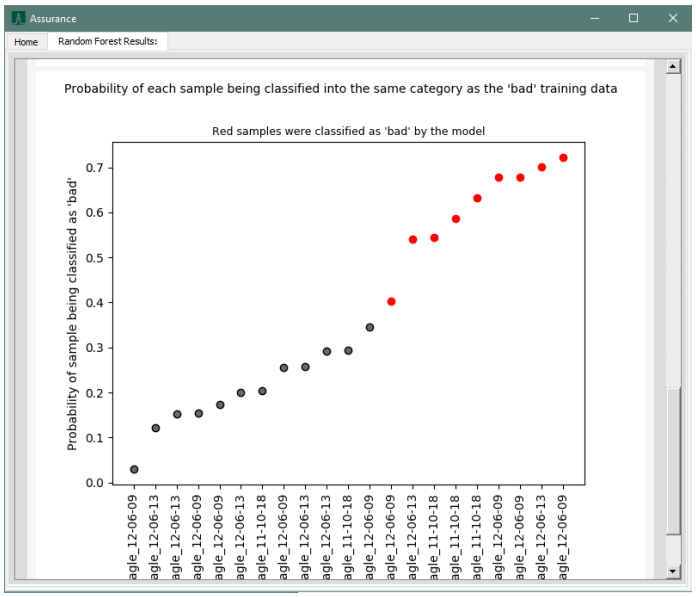


Figure 4.11 - Screenshot of the proportion of trees that voted each sample as 'bad' if a graph of the identification data was used.

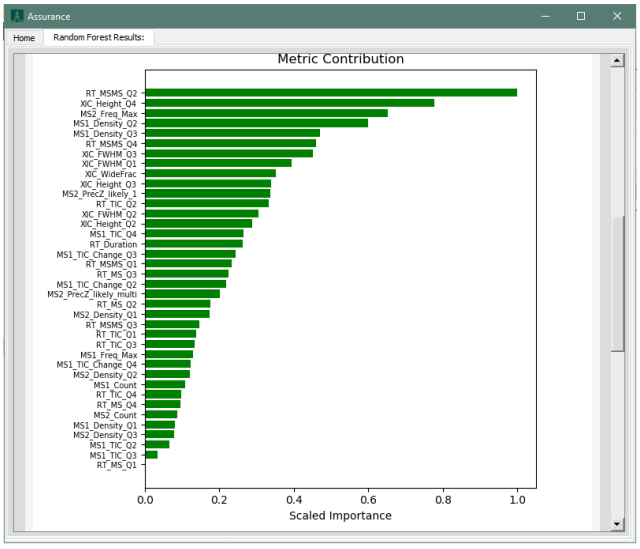


Figure 4.12: Screenshot of the metric contribution for the random forest analysis via table of the quality metrics.

Working from the assumption that the manual curation performed as part of the original study was correct, classification via identification samples resulted in 13 correct predictions out of the

21 samples and classification via quality table resulted in 14 correct predictions (Table 4.2). The prediction from identification samples resulted in a ‘good’ predictive value of 0.46 and a ‘bad’ predictive value of 0.17, whereas the quality metrics based classification resulted in 0.4 and 0.69 respectively. However, the sensitivity was 0.5 and 0.28 and the specificity was 0.50 and 0.79 for the identification approach and the quality approach respectively (See Suppl. Table 3 and 4 for the confusion matrices). The quality based approach was therefore more specific and resulted in a more stringent classification of samples as ‘bad’ quality, however the identification classification approach resulted in higher correct predictions overall.

The manual curation made by experts in Amidan et al.,(2014) which consisted of a three level classification system, “good”, “ok” and “poor”, was made from “base peak chromatogram, total ion current chromatogram, plots of both the top 50 000 and top 500 000 LC−MS detected features, and the number of peptides identified”. The identification based strategy showing a higher correct prediction in this example, may therefore be related to the curation strategy.

Table 4.3 - Confusion matrix from the analysis made with a table of quality files

	Predicted ‘Good’	**Predicted ‘Bad’**	**Total**
Manually curated ‘ok’ and ‘good’	TP = 2	FN=5	**7**
Manually curated ‘poor’	FP =3	TN=11	**14**
Total	**5**	**16**	

Table 4.4 - Confusion matrix from the analysis where the training set was classified from a graph of identification files

	Predicted 'Good'	**Predicted 'Bad'**	**Total**
Manually curated 'ok' and 'good'	TP = 6	FN = 1	**7**
Manually curated 'poor'	FP = 6	TN = 8	**14**
Total	**12**	**9**	

4.4 Conclusion

Assurance provides a useful analysis tool for running command line quality tools, QuaMeter and SwaMe and analyzing data. It is clear that in the outlier detection method, two of the anomalies that were most prominent were such a great distance away from the rest of the data that the subsequent PCA was influenced by the outliers. This analysis therefore indicates the importance of applying insight to statistical results. Random forest analysis where classification was done via identification results was more sensitive and showed more correct predictions in this dataset, but a quality table approach was more specific.

Chapter 5: Discuss

5.1 QC and reproducibility in an identification-driven field

Traditionally the success of a proteome analysis technique, software or experiment is often measured by the number of peptide/protein identifications with the ultimate goal of observing as much of the proteome as possible.[200] It is certainly true that a publication bias toward higher identifications may drive the experimental focus away from reproducibility.[201] The question arises whether our competition for identifying the highest number of peptides/proteins is the best approach to enhance the proteomic field. In 2014, a retrospective analysis was conducted on two publications in the journal *Nature.* Both studies claimed to have identified the largest number of proteins in the human proteome to date. However, despite neither of the two studies including nasal tissue, 108 and 200 olfactory receptors were found in the two studies respectively.[202] At the time of writing this thesis, these publications had 650 and 551 citations respectively, indicating that despite the perception of uncontrolled false identifications, the proteomics community still considered these publications valid and important.

Although a study identifying many peptides can correlate with all the correct quality practices and analyses, the pursuit of reproducibility and high identification counts can be counterproductive. For example, by performing a more constrained search rather than seeking to identify as many different peptides as possible, the reproducibility of the analysis could be increased as is demonstrated in the recent unveiling of a new dual-search technique.[203]

In addition, due to budget constraints, a researcher may need to decide between replicates or fractions of the sample. Where multiple fractions are traditionally employed to lessen the complexity at the detector, thereby possibly increasing the number of identifications, replication is traditionally employed to increase reproducibility and statistical power. This choice may therefore ask a researcher to choose between reproducibility or the number of identifications.

The pursuit of high levels of reproducibility goes further than the quality metrics and experimental design discussed in this thesis. A 2016 study found 20-50% of false positive peptide identifications in the datasets they analysed to be the result of modified peptides.[204] As anyone in the field of science may know, irreproducible science is at its best unproductive; at its worst, it may be the spark for time-consuming and financially costly unproductive follow-up studies. Similarly destructive, an irreproducible study could mask significant and valuable results and without the proper study limitations noted, other researchers could be discouraged from pursuing valuable hypotheses. An argument can therefore be made that a shift in our scientific goals is needed, from a pursuit of achieving the highest number of identifications, to the highest degree of reproducibility.

The topic of reproducibility also brings about a discussion of metrics of reproducibility. Many articles utilize the overlap in peptide identifications as reproducibility of identification, without including quantitation. The question then arises whether the overlap in identification is sufficient as a metric by which to measure reproducibility. In the light of this thesis and the corresponding quality studies, one is tempted to ask whether quality metrics may prove a less biased assessment of reproducibility.

5.2 Overall study outcomes

There is a need for an in-depth QC metrics producing software for DIA. Furthermore, the possibility that a bench biochemist can perform all the rest of the analysis without learning a statistical language, provides a clear motive for increasing the accessibility of command line quality tools and their downstream analysis. The present study involved the creation of SwaMe, a metrics producing software for DIA and Assurance, a statistical downstream analysis tool for analysing quality metrics. The overall aim of the study was to provide researchers with the necessary tools as well as to demonstrate the quality analysis of proteomics.

Much emphasis was placed in this study on identifying sources of variability. This step is extremely important both in method development, but also in verifying that there were no batch effects present in a study and that the experimental design was sufficient at the conclusion of the experiment. If fractionation can be taken as an example, the large contribution that this step has to the experimental variability should be considered in light of the evidence that the number of identifications of a sample were not always increased as we showed in the DDA analysis in chapter two. In addition, studies in DIA have found that libraries generated from fractionated DDA samples did not always improve the number of DIA identifications.[206] Such analysis could just as easily be used in common quality control analysis to find areas of the analysis where analysts can improve the reproducibility of their technique.

Chapter three on the other hand highlighted that for DIA, the isolation window structure plays a large role in the quality metrics. The experimental conditions, outside of LCMS analysis, such as cell culture were also highlighted as a major contributor to sample variability. Although it may have been expected in the peptide identifications, it was interesting to see this trend in the quality metrics. Even in chromatography based metrics, this trend was visible indicating that it is not simply a case of the number of peptides being comparable between groups. I hypothesize that this tendency is rather due to the different peptides being produced during the different growth phases. The similarity between the last two groups also suggest that the organism may have been entering a stationary phase or at least that the exponential growth phase was starting to plateau.

With both chapters two and three demonstrating how to inspect a dataset for variability, a researcher with access to a statistical language such as R or another statistical tool can replicate the analysis in their own dataset. The degree of personalization (dividing the experiment into groups) required to conduct this type of analysis is not currently possible with Assurance, but is definitely planned as a future feature.

The datasets included in this study showed that some of the fundamental experimental design strategies such as blocking, randomization and random block design need to be emphasized in proteomics. Over the last two decades, there has been an increase in significance of experimental design in the proteomics community, perhaps due to a large amount of data being produced in the 1990's that could not be biologically validated.[207] It is not uncommon to observe a course on experimental design included in proteomics workshops and conferences,[208,209] and since the first journal article established this principle in 2004,[210] journals frequently offer guides on the implementation of such strategies.

A plethora of articles also exist on the subject to increase the visibility. Hu and colleagues discussed three cautionary case studies with much the same message as ours in chapter two.[211] Their cases cluster according to the run date rather than any of the biological trends, they reported biological groups that were run a month before the other groups indicating clear batch effects. Furthermore, they show a case study where the protocol was changed halfway through and due to a lack of Fisher's design principles this also resulted in a batch effect. Their third case study demonstrated instrumental changes resulting in a mass error, which was only applicable to a subgroup of their study which was incorrectly interpreted as a biological difference rather than as a poor design artifact. The question then arises, given the amount of data available to researchers that emphasize design, why did less than a fifth of the datasets in chapter two apply experimental design? By comparison, in chapter three, two out of the five datasets analyzed with SwaMe, the Stoychev dataset and the Steen dataset, did indeed apply randomized block design. I hypothesize that two factors could be responsible for the low percentage of datasets that used blocking/randomization, a trend that is especially prevalent in chapter two. Firstly, the date of the experiments plays a large role. Although none of the datasets were run prior to the Hu and colleagues article,[211] the argument can be made that the more recent studies have had more exposure to the principle of experimental design and that over time this issue may be corrected. This could explain the slightly higher percentage of

datasets with randomized and/or blocking design in chapter three as DIA is a more recent technique. Some of the chapter two experiments predate the popularization of this technique for proteomics.[8]

The second possibility that I hypothesize is that researchers may be employing experimental design in their bench practices, but as many biologists may send their samples to a core facility, it is possible that the design is not communicated to the core facility and that the run order is the only part of the experiment that does not follow the design principle. However, as mentioned in the study by Hu and colleagues, as well as with some of the studies in chapter two, one biological group is sometimes run weeks before the other, indicating the researcher is aware of the timeline. One of the metrics that is produced by both QuaMeter and SwaMe (and the first metric displayed by Assurance when working from data from these tools) is StartTimeStamp. As seen in chapter four, the 23-day gap in the run dates of the Tabb dataset is unmistakable and to get to any other metric, the researcher is first confronted with this gap in the timeline. If the researcher had sent samples away to be analyzed elsewhere, any ignorance in the experimental design is corrected.

An important consequence of a study such as this might be termed the “QC fail dilemma.” In this scenario, a researcher who has already completed an experiment and is, for example, writing up a manuscript may attempt a run with a quality control tool such as SwaMe and/or Assurance only to find one or more runs registering as anomalies. This researcher is now faced with a number of choices, depending on the experimental design of the experiment. If the study had ample funding for replicates, the researcher may be lucky enough that the outliers may have fallen in replicates of different samples and the number of replicates of said samples could simply be reduced. If no replicates are present, but the experimental design has allowed the outlier samples to occur in different conditions, assuming that rerunning the samples is not an option, this fact could simply be noted. However, in the worst case scenario, where multiple anomalies fall in the same condition/timeslot in a comparison study, this fact should be noted

very clearly in the manuscript, to avoid deceiving the readers. In this manner, anyone who reads the paper and may have conflicting data, or may be planning a similar study of their own, would know that a possibility exists that the trends seen in the data could be the result of a batch effect. If a researcher of a proteomics experiment is able to rerun samples, it is advised to rerun an entire block of samples again. In this manner, the samples within the block can be compared to each other and the time that has passed from the first runs to the second group will not result in a batch effect.

It is important to note that this study was focused mainly on proteomics with data-dependent or data-independent acquisition. Targeted acquisition has very different quality considerations to discovery, similarly, the acquisition strategy impacts the dynamic range, MS2 complexity and other factors.[212]

It is particularly gratifying from a traditional QC analysis viewpoint to note that SwaMe was able to detect the quality anomaly in the Steen dataset prior to the occurrence. A mass spectrometer is made up of connected parts, not all of which can be monitored in real time. It is therefore very important to be able to narrow down the search to a specific type of problem as manufacturers guidelines indicate the problem could be occurring in any of a number of steps.[213–215]

If the different outlier detection techniques used throughout this study are compared, we are presented with interesting results. The robust PCA, identification analysis and non-robust PCA identified the same two samples as anomalies in the Tabb dataset, SW2-1-9 and SW2-1-10. However, the additional two samples that were detected in the identification method indicate some of the fundamental differences between using an identification method and a quality based one. It is still common conduct to use the number of identified peptides as a measure of the quality in a discovery experiment,[216] or at least combined with quality metrics.[168] It is therefore worthwhile to note the runs that show a difference in these metrics. In the case of the Tabb dataset, three out of the four of the fraction number 10 runs from the SAWC3561 sample were shown outliers by the identification data. Although we know that SAWC3561-2-1-10 is an

aberration in quality, with so much data at hand it would appear that perhaps the other three are possibly a biological trend. The 10th fraction was the last, so a slight difference in the fractionation process which consisted of a manual selection of gel segments could have resulted in less peptides in this fraction. This is still true despite the fact that we have complimented the analyst on a remarkably reproducible gel segment selection in the experiment as a whole. Several additional samples were selected in chapter four where a non-robust PCA was used. The influence of anomalies on classical PCA is well-known in statistics.[217]

In addition, Tukey's criteria for an outlier always assumes a certain degree of symmetry, and when anomalies as far from the rest of the data as those in the Tabb dataset occur, the skewness that results may highlight a larger group of data points as outliers than expected.[218] It is therefore advisable to have Tukey's method for outlier detection paired with a more robust PCA. However, the dataset-specific characteristics of a robust PCA made it a little more difficult to implement in the generalized setting of Assurance. This indicates that, in Assurance or any other setting where a classical PCA is used, a certain degree of manual interpretation is necessary in cases of extreme aberrations to ascertain whether all the proposed ProbOuts are indeed probable outliers or whether some of the anomalies present have skewed the data.

All of the above mentioned methods are based on PCA. However, both semi-supervised and supervised methods have been investigated for mass spectrometry quality control.[114,169] Supervised learning has the value of allowing expertly curated data as a classifier. In addition, where historical data are available, problem runs can be flagged as patterns to seek in future. Unlike unsupervised methods which have a garbage in / garbage out problem, supervised classification can classify problem areas even if the entire dataset falls under the 'bad' quality banner. However, there are a number of drawbacks to using supervised methods such as the random forest approach in Assurance (reviewed by Su, 2011[219]). Firstly, it can be hard to correctly classify data in a rapidly changing environment and therefore constant re-assessment of the training data is necessary. In addition, the technique is extremely reliant on correctly

allocated training data and even slight deviances in the training data can throw the entire model off. Lastly, it is possible that a different type of problem from the problems added to the training set occur in the dataset for classification and is then missed by the supervised classification as this problem may closer resemble other data points in the 'good' quality group than the 'bad' quality group as it pertains to changes in different sets of metrics. It is therefore important to have the training set encompass as many different problem conditions as possible.

5.3 Significance of the study

Applications of discovery proteomics include biomarker discovery research such as the search for diagnostic and treatment response biomarkers, drug discovery and identifying disease mechanisms. These medical applications are paramount in the fight against both communicable and noncommunicable diseases,[220–223] and their importance in the present COVID-19 pandemic has also been illustrated.[224–226]

Dr. Rodriguez also shared HUPO-PSI's vision for unified metrics, in the pursuit of a more reproducible system.[227] HUPO-PSI QC team aims to address this issue in the form of the controlled vocabulary of mzQC as discussed in chapter one. 2020 has illuminated a situation where universal metrics are sorely needed. During the pandemic, researchers found great value in preprints/ articles that have not yet been peer-reviewed in fighting an active epi/pandemic. With universal quality metrics, the ability to judge for oneself the validity of a study before peer-review becomes a less daunting task. For example, the QuaMeter metric "StartTimeStamp" portrayed in a graph as in chapter 1,2 and 3 can immediately show whether a randomized, block or randomized block design was conducted in the runorder. Experimental design is such a general scientific term that a genetics researcher for example should be able to spot a batch effect whilst reading a proteomics study.[228,229] In fact, batch effects in genomics studies sometimes even reach popular media and are read by non-scientists.[230] Such situations are embarrassing for journal and researcher alike (or at least should be), and I have no doubt that

universal metrics such as those produced by QuaMeter or SwaMe will become commonplace within the next decade.

As the world of proteomic quality advances, bioinformatic software must be ready for these changes. SwaMe writes and Assurance reads mzQC files, the new format in process from HUPO-PSI.[227] By implementing this new file type and incorporating SwaMe's metrics in the controlled vocabulary, the software usage is expanded. In addition, an important part of the design of both software tools was to make the code available to all and share the knowledge of the metric calculations. With such knowledge, the metrics can be improved upon by others in the scientific community as the instrument usage grows and changes. The collaboration with the University of Manchester in the production of SwaMe has also broadened the scope of SwaMe's implementation and allowed fresh insight and an outlook and interpretation unique from our own. Our collaborators have devised a very intricate and informative set of metrics that come especially handy in longitudinal analysis. For example, they have thought of using the stability awarded by calibrant peptides in the quality setting and have included metrics such as iRTOrderedness. Together, SwaMe has become a more powerful, useful tool that will be more robust to changes to the technique in the future.

SwaMe is the first software of its kind for DIA proteomics. The tool allows an in-depth analysis to a point that has never been possible before by segmenting the RT and m/z axis in the form of different windows and providing metrics for each segment. DIA is marketed as more reproducible than DDA proteomics,[206,231,232] which leaves the concern that the statement may be interpreted by researchers as an excuse to not implement proper QC protocols. It is therefore important to improve accessibility of QC in this field so that this factor does not add to possible reasons proper QC is avoided.

This thesis relied heavily on public data repositories. The advent and rising popularity of data repositories enables not only the *a posteriori* quality analysis of data with quality tools,[200] the demonstration of novel uses for quality metrics such as those introduced in chapter two, but

also of course brings us closer to the goal of transparency.[205] This scale of data repository usage was not present a few decades ago and this study demonstrates the importance of public repositories.

5.4 Study limitations

Although care was taken to include instruments from three different vendors in our DIA analysis, regrettably, the new and exciting technology of TIMSTOF and associated ion mobility (in-source or in mass spectrometer) were not included in our design of SwaMe. It is very possible that the popularity of these techniques might increase in the future, rendering their absence from SwaMe a limitation of the software.

Assurance is a windows-based software, however, this decision was made with the reasoning that most command-line/programming shy users would also not utilize Linux.

SwaMe takes as input the standard HUPO file format, .mzML. There are analysis tools that do not require the conversion from raw file to standard format,[24] however some analysis tools that do not require this format advise the conversion for optimal performance,[103] and there are others that cannot operate without the conversion.[194,195,233]

As mentioned previously, Assurance currently is not equipped to show different grouping structure in different colours in a PCA graph. This is due to time constraints, but would be very interesting to add at a later stage.

In addition, SwaMe does not consider identification-based metrics. As there is such a discrepancy between different identification software, this avenue could be considered a limitation and could be useful for someone trying to compare bioinformatic methods.

Chapter 6: Conclusion and future works

6.1 Conclusion

The aims of this study included creating tools for the analysis of DIA proteomic data QC metric generation and the downstream analysis of metrics for both DDA and DIA. In addition, I set out to illustrate how quality metrics could be used in a conventional manner to detect outliers as well as a novel probing of quality metrics to determine the main sources of variability in the rest of the experiment.

In addition, I highlighted the importance of experimental design and how quality tools can be used to ascertain whether the correct experimental design strategy was chosen. In chapter 5, the significance of the study and the contribution to the field was discussed along with the study limitations.

This thesis has produced one published article and two publication ready manuscripts. I have presented at one national proteomics workshop hosted in South Africa,[234] one national bioinformatics conference,[235] as well as hosting a tutorial for the detection of quality outliers in mass spectrometry data at an international meeting of HUPO-PSI in 2019,[236] hosted in Cape Town. At this meeting the collaboration between University of Manchester and Stellenbosch University began after realizing that we had common goals, but orthogonal objectives that could work together very well. It was also at the workshop I presented at this meeting that the rationale for creating Assurance became clear as a quick survey eluded that the researchers attending the workshop had mostly either never used any statistical toolkit/language such as R/ python or they refrain from using such tools wherever possible.

6.2 Future works

I believe that Assurance is a project that has much potential to grow further. Possible features that could be added include the ability to differentiate groups in any of the graphs. For some metrics, more informative graph styles might help visualize those metrics (for example a stacked bar chart for RT-TIC metrics). In addition, a more detailed description of what each metric stands for and how it could be interpreted could be added to SwaMe and QuaMeter metrics.

Unsupervised and supervised outlier detection methods were selected, however there are many other methods available that could be incorporated as options, such as isolation forest[237] and factor analysis.[238] One could allow the parameters of the random forest amongst other methods to be set by the user. In addition, the method of declaring an outlier can become an option between z-score, modified z-score and MAD^e (see review[218]).

There is also room for new metrics to be added to SwaMe. The decision to exclude identification in SwaMe analysis was taken to prevent identification abnormalities masking instrument anomalies. However, the gap between the identification-based outlier detection of the Stoychev team and the identification-independent outlier detection performed with SwaMe highlighted some of the value of including identification results in quality analysis. Including an identification mode in SwaMe in future, could therefore be considered.

The public availability of the code, (SwaMe:https://github.com/PaulBrack/Yamato, Assurance: https://github.com/marinaPauw/Assurance) as well as the collaborative nature of the project will hopefully enable and increase the probability that the project will be worked on and added to in the future. There is particularly room for addition of metrics related to the m/z axis. A metric such as resolution could for example be added to make this area stronger. In addition the m/z that is most commonly a base peak of a scan could be pointed out in a metric.

SwaMe could also benefit from an identification-based mode. If this avenue is taken, one might even go further to allow the analysis of for example Geyer and colleagues's panel of plasma quality index proteins.[23] In the discussion of their article, it was noted that the panel could be

adapted for other sample types and once other panels have been developed, they too could find a place in an identification-based quality software.

SwaMe also focused on specific implementations of DIA, hence a technique like ion mobility for example is not accurately represented in the metrics. Incorporating ion mobility features would greatly broaden the set of MS^E data sets on which SwaMe could be usefully deployed.

Recent studies have also found success in real-time database searching. [239] The idea revolves around performing a database search while the instrument is still gathering data, even deciding which precursors to fragment on the basis of those identified so far. These advances, particularly when added to the speed of spectral library-based search, enable the rapid production of identification-based metrics. If such a technology is incorporated into Assurance, it could open up an entire world of combining quality tools with identification in one platform. Going the real-time route with quality control software will have tremendous value.

It may also be useful to conduct a meta-analysis of all datasets submitted to a repository such as PRIDE in the last year and perform an anonymous analysis of experimental design to obtain a global figure of how many studies implemented Fisher's design principles.

There is therefore room for growth both in the analysis and in the development of software parts of this study.

6.3 Proteomics QC in SA

In a broader sense, proteomics in SA has ample room for improvement. Many instruments are located hundreds of kilometers from their nearest neighbours, and an active virtual community could make the operation and quality control of these instruments much easier. I suggest this could be handled by implementing a similar strategy to one adopted in the National Laboratory Association (NLA).[240] In the field of microbiology, the association sends out samples of which the composition/cfu count is known to the association, but not the analysts. Upon sending results back, the laboratory can compare its unique and anonymous number to that of other

laboratories in the country. This would be similar to analyzing an *E.coli* digest or other complex sample as a QC step, but it not only allows the comparison of the instrument with a previous version, but also allows a researcher to compare their instrument and techniques to other laboratories in the country. This is important as standards of acceptable levels of variability within an instrument QC metrics can differ with the personality of the analyst, however, feelings of inadequacy towards fellow institutions may put things into perspective and spark action.

Notably, due to a relatively low rate of exchange for our currency and the location of our country, attending an international conference or vendor training is a tremendous financial expense for South Africans.[241] A meeting such as the 2018 HUPO-PSI meeting where international scientists visit our country allows many South African scientists to converse and collaborate with international scientists, creating relationships and starting projects together. My project has shown that this type of international meeting can result in fruitful collaborations and potential articles, as well as allowing scientists to discuss solutions to possible problems.

In addition, our country would profit from an annual meeting for mass spectrometry specialists and students. The South African Department of Science and Technology initiative, Diplomics,[242] and the African Centre for Gene Technologies, ACGT,[243] have been increasingly active in South Africa in organizing international workshops such as the Advanced proteomics course in 2018 and the Skyline and Trans-proteomic Pipeline proteomics courses in 2019. These courses work to increase the skill-level within South Africa without the costly expense of sending scientists abroad. However, these courses are typically attended by students, whereas a proteomics meeting might include both students and laboratory heads and allow presentation of specifically proteomics projects.

The pandemic of 2020 presented many challenges for academic conferences, however, it is my hope that the virtual conferences of this year may consider maintaining some of their virtual ability in the years to come. A virtual conference aids in breaking down barriers for lower income country students to attend. By foregoing transport costs, an entire laboratory in South Africa

could now attend a virtual conference hosted overseas, where before the laboratory might only be able to send one candidate at most in a year. An increase in remote technologies in all forms, conferences, technological support etc. will help South African proteomics to grow.

6.4 Concluding remarks

In the previous chapter, the study limitations were discussed, but in this project, as with so many others, the main hindrance was time. There is so much more that could be added to both softwares to grow their feature sets and improve their usability. It is my hope that the tools created as well as the information illuminated by this study will not only be used, but also be improved upon in the future.

Chapter 7: References

(1) Makarov, A.; Denisov, E.; Lange, O.; Horning, S. Dynamic Range of Mass Accuracy in LTQ Orbitrap Hybrid Mass Spectrometer. *J. Am. Soc. Mass Spectrom.* **2006**, *17* (7), 977–982. https://doi.org/10.1016/j.jasms.2006.03.006.

(2) Glish, G. L.; Burinsky, D. J. Hybrid Mass Spectrometers for Tandem Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2008**, *19* (2), 161–172. https://doi.org/10.1016/j.jasms.2007.11.013.

(3) Zubarev, R. A.; Makarov, A. Orbitrap Mass Spectrometry. *Anal. Chem.* **2013**, *85* (11), 5288–5296. https://doi.org/10.1021/ac4001223.

(4) Liu, H.; Lin, D.; Yates, J. R. Multidimensional Separations for Protein/Peptide Analysis in the Post-Genomic Era. *BioTechniques* **2002**, *32* (4), 898–911. https://doi.org/10.2144/02324pt01.

(5) Yates, J. R. Mass Spectrometry and the Age of the Proteome. *J. MASS Spectrom.* **1998**, *33*, 19.

(6) Eng, J. K.; Searle, B. C.; Clauser, K. R.; Tabb, D. L. A Face in the Crowd: Recognizing Peptides Through Database Search. *Mol. Cell. Proteomics* **2011**, *10* (11). https://doi.org/10.1074/mcp.R111.009522.

(7) Michalski, A.; Cox, J.; Mann, M. More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority Is Inaccessible to Data-Dependent LC−MS/MS. *J. Proteome Res.* **2011**, *10* (4), 1785–1793. https://doi.org/10.1021/pr101060v.

(8) Gillet, L. C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. Targeted Data Extraction of the MS/MS Spectra Generated by Data-Independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Mol. Cell. Proteomics* **2012**, *11* (6), O111.016717. https://doi.org/10.1074/mcp.O111.016717.

(9) Tsou, C.-C.; Avtonomov, D.; Larsen, B.; Tucholska, M.; Choi, H.; Gingras, A.-C.; Nesvizhskii, A. I. DIA-Umpire: Comprehensive Computational Framework for Data-Independent Acquisition Proteomics. *Nat. Methods* **2015**, *12* (3), 258–264. https://doi.org/10.1038/nmeth.3255.

(10) Guan, S.; Taylor, P. P.; Han, Z.; Moran, M. F.; Ma, B. Data Dependent–Independent Acquisition (DDIA) Proteomics. *J. Proteome Res.* **2020**. https://doi.org/10.1021/acs.jproteome.0c00186.

(11) Whitney, C. W.; Lind, B. K.; Wahl, P. W. Quality Assurance and Quality Control in Longitudinal Studies. *Epidemiol. Rev.* **1998**, *20* (1), 71–80. https://doi.org/10.1093/oxfordjournals.epirev.a017973.

(12) Taylor, C. F.; Hermjakob, H.; Julian, R. K.; Garavelli, J. S.; Aebersold, R.; Apweiler, R. The Work of the Human Proteome Organisation's Proteomics Standards Initiative (HUPO PSI). *Omics J. Integr. Biol.* **2006**, *10* (2), 145–151. https://doi.org/10.1089/omi.2006.10.145.

(13) Rudnick, P. A.; Clauser, K. R.; Kilpatrick, L. E.; Tchekhovskoi, D. V.; Neta, P.; Bunk, D. M.; Cardasis, H. L.; Ham, A.-J. L.; Jaffe, J. D.; Kinsinger, C. R.; Mesri, M.; Neubert, T. A.; Schilling, B.; Tabb, D. L.; Tegeler, T. J.; Vega-Montoto, L.; Variyath, A. M.; Wang, M.; Wang, P.; Whiteaker, J. R.; Zimmerman, L. J.; Carr, S. A.; Fisher, S. J.; Gibson, B. W.; Paulovich, A. G.; Regnier, F. E.; Rodriguez, H.; Spiegelman, C.; Tempst, P.; Liebler, D. C.; Stein, S. E. Performance Metrics for Liquid Chromatography-Tandem Mass Spectrometry Systems in Proteomics Analyses*□S. 17.

(14) U.S. Pharmacopeia https://www.usp.org/ (accessed Jun 16, 2020).

(15) Guidance for Industry #118 - Mass Spectrometry for Confirmation of the Identity of Animal Drug Residues - Final Guidance, May 1, https://webcache.googleusercontent.com/search?q=cache:IjUi-j8EFcsJ:https://www.fda.gov/media/70154/download+&cd=2&hl=en&ct=clnk&gl=za (accessed May 17, 2020).
(16) Bernstein, S. Sur l'extension du théoréme limite du calcul des probabilités aux sommes de quantités dépendantes. *Math. Ann.* **1927**, *97* (1), 1–59. https://doi.org/10.1007/BF01447859.
(17) Oberg, A. L.; Vitek, O. Statistical Design of Quantitative Mass Spectrometry-Based Proteomic Experiments. *J. Proteome Res.* **2009**, *8* (5), 2144–2156. https://doi.org/10.1021/pr8010099.
(18) Altman, D. G.; Bland, J. M. How to Randomise. *BMJ* **1999**, *319* (7211), 703–704.
(19) Bang, J. W. An Empirical Comparison of Random Number Generators: Period, Structure, Correlation, Density, and Efficiency. PhD Thesis, University of North Texas, Denton, TX, USA, 1995.
(20) Rotz, W.; Joshee, A.; Falk, E. A Comparison of Random Number Generators Used in Business - 2004 Update. 4.
(21) Chavez, I. Handbook 133 - 2019 (Current Edition) https://www.nist.gov/pml/weights-and-measures/publications/nist-handbooks/other-nist-handbooks/other-nist-handbooks-2 (accessed Nov 7, 2019).
(22) Hassis, M. E.; Niles, R. K.; Braten, M. N.; Albertolle, M. E.; Ewa Witkowska, H.; Hubel, C. A.; Fisher, S. J.; Williams, K. E. Evaluating the Effects of Preanalytical Variables on the Stability of the Human Plasma Proteome. *Anal. Biochem.* **2015**, *478*, 14–22. https://doi.org/10.1016/j.ab.2015.03.003.
(23) Geyer, P. E.; Voytik, E.; Treit, P. V.; Doll, S.; Kleinhempel, A.; Niu, L.; Müller, J. B.; Buchholtz, M.-L.; Bader, J. M.; Teupser, D.; Holdt, L. M.; Mann, M. Plasma Proteome Profiling to Detect and Avoid Sample-Related Biases in Biomarker Studies. *EMBO Mol. Med.* **2019**, *11* (11), e10427. https://doi.org/10.15252/emmm.201910427.
(24) Tyanova, S.; Temu, Tikira; Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics | Nature Protocols https://www.nature.com/articles/nprot.2016.136 (accessed Apr 21, 2020).
(25) O'Mullan, P.; Craft, D.; Yi, J.; Gelfand, C. A. Thrombin Induces Broad Spectrum Proteolysis in Human Serum Samples. *Clin. Chem. Lab. Med. CCLM* **2009**, *47* (6), 685–693. https://doi.org/10.1515/CCLM.2009.003.
(26) Percy, A. J.; Parker, C. E.; Borchers, C. H. Pre-Analytical and Analytical Variability in Absolute Quantitative MRM-Based Plasma Proteomic Studies. *Bioanalysis* **2013**, *5* (22), 2837–2856. https://doi.org/10.4155/bio.13.245.
(27) Omenn, G. S. The Human Proteome Organization Plasma Proteome Project Pilot Phase: Reference Specimens, Technology Platform Comparisons, and Standardized Data Submissions and Analyses. *PROTEOMICS* **2004**, *4* (5), 1235–1240. https://doi.org/10.1002/pmic.200300686.
(28) Loo, R. R.; Dales, N.; Andrews, P. C. Surfactant Effects on Protein Structure Examined by Electrospray Ionization Mass Spectrometry. *Protein Sci. Publ. Protein Soc.* **1994**, *3* (11), 1975–1983.
(29) Tabb, D. L.; Murugan, B. D.; Okendo, J.; Nair, O.; Blackburn, J. M.; Buthelezi, S. G.; Stoychev, S. Open Search Unveils Modification Patterns in Formalin-Fixed, Paraffin-Embedded Thermo HCD and SCIEX TripleTOF Shotgun Proteomes. *Int. J. Mass Spectrom.* **2020**, *448*, 116266. https://doi.org/10.1016/j.ijms.2019.116266.
(30) Proteomic Analysis of Formalin-fixed Prostate Cancer Tissue | Molecular & Cellular Proteomics https://www.mcponline.org/content/4/11/1741.full (accessed May 21, 2020).
(31) Magdeldin, S.; Yamamoto, T. Toward Deciphering Proteomes of Formalin-Fixed Paraffin-

Embedded (FFPE) Tissues. *Proteomics* **2012**, *12* (7), 1045–1058. https://doi.org/10.1002/pmic.201100550.
(32) Shi, S.-R.; Liu, C.; Balgley, B. M.; Lee, C.; Taylor, C. R. Protein Extraction from Formalin-Fixed, Paraffin-Embedded Tissue Sections: Quality Evaluation by Mass Spectrometry. *J. Histochem. Cytochem. Off. J. Histochem. Soc.* **2006**, *54* (6), 739–743. https://doi.org/10.1369/jhc.5B6851.2006.
(33) Ericsson, C.; Nistér, M. Protein Extraction from Solid Tissue. In *Methods in Biobanking*; Dillner, J., Ed.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2011; Vol. 675, pp 307–312. https://doi.org/10.1007/978-1-59745-423-0_17.
(34) Ericsson, C.; Peredo, I.; Nistér, M. Optimized Protein Extraction from Cryopreserved Brain Tissue Samples. *Acta Oncol.* **2007**, *46* (1), 10–20. https://doi.org/10.1080/02841860600847061.
(35) Jiang, X.; Jiang, X.; Feng, S.; Tian, R.; Ye, M.; Zou, H. Development of Efficient Protein Extraction Methods for Shotgun Proteome Analysis of Formalin-Fixed Tissues. *J. Proteome Res.* **2007**, *6* (3), 1038–1047. https://doi.org/10.1021/pr0605318.
(36) Daniel, C.; Triboï, E. Isolation of Wheat Grain Compartments and Their Protein Composition. *Cereal Res. Commun.* **2001**, *29* (1–2), 197–204. https://doi.org/10.1007/BF03543661.
(37) Tan, S.; Tan, H. T.; Chung, M. C. M. Membrane Proteins and Membrane Proteomics. *PROTEOMICS* **2008**, *8* (19), 3924–3932. https://doi.org/10.1002/pmic.200800597.
(38) Cox, B.; Emili, A. Tissue Subcellular Fractionation and Protein Extraction for Use in Mass-Spectrometry-Based Proteomics. *Nat. Protoc.* **2006**, *1* (4), 1872-.
(39) Stimpson, S. E.; Coorssen, J. R.; Myers, S. J. Optimal Isolation of Mitochondria for Proteomic Analyses. *Anal. Biochem.* **2015**, *475*, 1–3. https://doi.org/10.1016/j.ab.2015.01.005.
(40) Extraction of Membrane Proteins | SpringerLink https://link.springer.com/protocol/10.1385/1-59259-655-X:283 (accessed Jun 15, 2020).
(41) Cole, E. G.; Mecham, D. K. Cyanate Formation and Electrophoretic Behavior of Proteins in Gels Containing Urea. *Anal. Biochem.* **1966**, *14* (2), 215–222. https://doi.org/10.1016/0003-2697(66)90129-1.
(42) Manson, W. The Effect upon Casein of Aqueous Solutins of Urea. *Biochim. Biophys. Acta* **1962**, *63* (3), 515–517. https://doi.org/10.1016/0006-3002(62)90118-X.
(43) Cole, R. D. On the Transformation of Insulin in Concentrated Solutions of Urea. *J. Biol. Chem.* **1961**, *236* (10), 2670–2671.
(44) Boja, E. S.; Fales, H. M. Overalkylation of a Protein Digest with Iodoacetamide. *Anal. Chem.* **2001**, *73* (15), 3576–3582. https://doi.org/10.1021/ac0103423.
(45) Tsiatsiani, L.; Heck, A. J. R. Proteomics beyond Trypsin. *FEBS J.* **2015**, *282* (14), 2612–2626. https://doi.org/10.1111/febs.13287.
(46) Walmsley, S. J.; Rudnick, P. A.; Liang, Y.; Dong, Q.; Stein, S. E.; Nesvizhskii, A. I. Comprehensive Analysis of Protein Digestion Using Six Trypsins Reveals the Origin of Trypsin As a Significant Source of Variability in Proteomics. *J. Proteome Res.* **2013**, *12* (12), 5666–5680. https://doi.org/10.1021/pr400611h.
(47) Gong, J.-S.; Li, W.; Zhang, D.-D.; Xie, M.-F.; Yang, B.; Zhang, R.-X.; Li, H.; Lu, Z.-M.; Xu, Z.-H.; Shi, J.-S. Biochemical Characterization of An Arginine-Specific Alkaline Trypsin from Bacillus Licheniformis. *Int. J. Mol. Sci.* **2015**, *16* (12), 30061–30074. https://doi.org/10.3390/ijms161226200.
(48) Chelulei Cheison, S.; Brand, J.; Leeb, E.; Kulozik, U. Analysis of the Effect of Temperature Changes Combined with Different Alkaline PH on the β-Lactoglobulin Trypsin Hydrolysis Pattern Using MALDI-TOF-MS/MS. *J. Agric. Food Chem.* **2011**, *59* (5), 1572–1581. https://doi.org/10.1021/jf1039876.
(49) Hildonen, S.; Halvorsen, T. G.; Reubsaet, L. Why Less Is More When Generating Tryptic

Peptides in Bottom-up Proteomics. *PROTEOMICS* **2014**, *14* (17–18), 2031–2041. https://doi.org/10.1002/pmic.201300479.

(50) Loziuk, P. L.; Wang, J.; Li, Q.; Sederoff, R. R.; Chiang, V. L.; Muddiman, D. C. Understanding the Role of Proteolytic Digestion on Discovery and Targeted Proteomic Measurements Using Liquid Chromatography Tandem Mass Spectrometry and Design of Experiments. *J. Proteome Res.* **2013**, *12* (12), 5820–5829. https://doi.org/10.1021/pr4008442.

(51) Piehowski, P. D.; Petyuk, V. A.; Orton, D. J.; Xie, F.; Moore, R. J.; Ramirez-Restrepo, M.; Engel, A.; Lieberman, A. P.; Albin, R. L.; Camp, D. G.; Smith, R. D.; Myers, A. J. Sources of Technical Variability in Quantitative LC–MS Proteomics: Human Brain Tissue Sample Analysis. *J. Proteome Res.* **2013**, *12* (5), 2128–2137. https://doi.org/10.1021/pr301146m.

(52) Epstein, D. M.; Tebbett, I. R.; Boyd, S. E. Eliminating Sources of Pipetting Error in the Forensic Laboratory. *Forensic Sci. Commun.* **2003**, *5* (4), 6.

(53) Björhall, K.; Miliotis, T.; Davidsson, P. Comparison of Different Depletion Strategies for Improved Resolution in Proteomic Analysis of Human Serum Samples. *PROTEOMICS* **2005**, *5* (1), 307–317. https://doi.org/10.1002/pmic.200400900.

(54) Wang, G.; Wu, W. W.; Zeng, W.; Chou, C.-L.; Shen, R.-F. Label-Free Protein Quantification Using LC-Coupled Ion Trap or FT Mass Spectrometry: Reproducibility, Linearity, and Application with Complex Proteomes. *J. Proteome Res.* **2006**, *5* (5), 1214–1223. https://doi.org/10.1021/pr050406g.

(55) Whiteaker, J. R.; Zhang, H.; Eng, J. K.; Fang, R.; Piening, B. D.; Feng, L.-C.; Lorentzen, T. D.; Schoenherr, R. M.; Keane, J. F.; Holzman, T.; Fitzgibbon, M.; Lin; Zhang, H.; Cooke, K.; Liu, T.; Camp, D. G.; Anderson, L.; Watts, J.; Smith, R. D.; McIntosh, M. W.; Paulovich, A. G. Head-to-Head Comparison of Serum Fractionation Techniques. *J. Proteome Res.* **2007**, *6* (2), 828–836. https://doi.org/10.1021/pr0604920.

(56) Ichibangase, T.; Moriya, K.; Koike, K.; Imai, K. Limitation of Immunoaffinity Column for the Removal of Abundant Proteins from Plasma in Quantitative Plasma Proteomics. *Biomed. Chromatogr. BMC* **2009**, *23* (5), 480–487. https://doi.org/10.1002/bmc.1139.

(57) Gundry, R. L.; White, M. Y.; Nogee, J.; Tchernyshyov, I.; Van Eyk, J. E. Assessment of Albumin Removal from an Immunoaffinity Spin Column: Critical Implications for Proteomic Examination of the Albuminome and Albumin-Depleted Samples. *Proteomics* **2009**, *9* (7), 2021–2028. https://doi.org/10.1002/pmic.200800686.

(58) De Bock, M.; de Seny, D.; Meuwis, M.-A.; Servais, A.-C.; Minh, T. Q.; Closset, J.; Chapelle, J.-P.; Louis, E.; Malaise, M.; Merville, M.-P.; Fillet, M. Comparison of Three Methods for Fractionation and Enrichment of Low Molecular Weight Proteins for SELDI-TOF-MS Differential Analysis. *Talanta* **2010**, *82* (1), 245–254. https://doi.org/10.1016/j.talanta.2010.04.029.

(59) Hakimi, A.; Auluck, J.; Jones, G. D. D.; Ng, L. L.; Jones, D. J. L. Assessment of Reproducibility in Depletion and Enrichment Workflows for Plasma Proteomics Using Label-Free Quantitative Data-Independent LC-MS. *PROTEOMICS* **2014**, *14* (1), 4–13. https://doi.org/10.1002/pmic.201200563.

(60) Kirkwood, K. J.; Ahmad, Y.; Larance, M.; Lamond, A. I. Characterization of Native Protein Complexes and Protein Isoform Variation Using Size-Fractionation-Based Quantitative Proteomics. *Mol. Cell. Proteomics* **2013**, *12* (12), 3851–3873. https://doi.org/10.1074/mcp.M113.032367.

(61) Yeh, T.-T.; Ho, M.-Y.; Chen, W.-Y.; Hsu, Y.-C.; Ku, W.-C.; Tseng, H.-W.; Chen, S.-T.; Chen, S.-F. Comparison of Different Fractionation Strategies for In-Depth Phosphoproteomics by Liquid Chromatography Tandem Mass Spectrometry. *Anal. Bioanal. Chem.* **2019**, *411* (15), 3417–3424. https://doi.org/10.1007/s00216-019-01823-0.

(62) Lee, H.-J.; Lee, E.-Y.; Kwon, M.-S.; Paik, Y.-K. Biomarker Discovery from the Plasma Proteome Using Multidimensional Fractionation Proteomics. *Curr. Opin. Chem. Biol.* **2006**,

10 (1), 42–49. https://doi.org/10.1016/j.cbpa.2006.01.007.
(63) Scheerlinck, E.; Dhaenens, M.; Van Soom, A.; Peelman, L.; De Sutter, P.; Van Steendam, K.; Deforce, D. Minimizing Technical Variation during Sample Preparation Prior to Label-Free Quantitative Mass Spectrometry. *Anal. Biochem.* **2015**, *490*, 14–19. https://doi.org/10.1016/j.ab.2015.08.018.
(64) Larger, P. J.; Breda, M.; Fraier, D.; Hughes, H.; James, C. A. Ion-Suppression Effects in Liquid Chromatography–Tandem Mass Spectrometry Due to a Formulation Agent, a Case Study in Drug Discovery Bioanalysis. *J. Pharm. Biomed. Anal.* **2005**, *39* (1), 206–216. https://doi.org/10.1016/j.jpba.2005.03.009.
(65) Zhang, N.; Li, L. Effects of Common Surfactants on Protein Digestion and Matrix-Assisted Laser Desorption/Ionization Mass Spectrometric Analysis of the Digested Peptides Using Two-Layer Sample Preparation. *Rapid Commun. Mass Spectrom. RCM* **2004**, *18* (8), 889–896. https://doi.org/10.1002/rcm.1423.
(66) Palmblad, M.; Vogel, J. Quantitation of Binding, Recovery and Desalting Efficiency of Peptides and Proteins in Solid Phase Extraction Micropipette Tips. *J. Chromatogr. B* **2005**, *814* (2), 309–313. https://doi.org/10.1016/j.jchromb.2004.10.052.
(67) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. Quantitative Analysis of Complex Protein Mixtures Using Isotope-Coded Affinity Tags. *Nat. Biotechnol.* **1999**, *17* (10), 994. https://doi.org/10.1038/13690.
(68) Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J. Multiplexed Protein Quantitation in Saccharomyces Cerevisiae Using Amine-Reactive Isobaric Tagging Reagents. *Mol. Cell. Proteomics* **2004**, *3* (12), 1154–1169. https://doi.org/10.1074/mcp.M400129-MCP200.
(69) Ong, S.-E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Mol. Cell. Proteomics* **2002**, *1* (5), 376–386. https://doi.org/10.1074/mcp.M200025-MCP200.
(70) Karp, N. A.; Huber, W.; Sadowski, P. G.; Charles, P. D.; Hester, S. V.; Lilley, K. S. Addressing Accuracy and Precision Issues in ITRAQ Quantitation. *Mol. Cell. Proteomics* **2010**, *9* (9), 1885–1897. https://doi.org/10.1074/mcp.M900628-MCP200.
(71) McCalley, D. V. Effect of Buffer on Peak Shape of Peptides in Reversed-Phase High Performance Liquid Chromatography. *J. Chromatogr. A* **2004**, *1038* (1–2), 77–84. https://doi.org/10.1016/j.chroma.2004.03.038.
(72) McCalley, D. V. Choice of Buffer for the Analysis of Basic Peptides in Reversed-Phase HPLC. 8.
(73) High-Sensitivity TFA-free LC-MS for Profiling Histones https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3517135/ (accessed Mar 31, 2020).
(74) Correlation between peak capacity and protein sequence coverage in proteomics analysis by liquid chromatography-mass spectrometry/mass spectrometry - ScienceDirect https://www.sciencedirect.com/science/article/abs/pii/S002196731000659X (accessed Mar 31, 2020).
(75) Effects of Column Length, Particle Size, Flow Rate, and Pressure Programming Rate on Resolution in Pressure-Programmed Supercritical Fluid Chromatography* | Journal of Chromatographic Science | Oxford Academic https://academic.oup.com/chromsci/article-abstract/18/2/75/287215?redirectedFrom=PDF (accessed Mar 31, 2020).
(76) Analysis of protein mixtures from whole-cell extracts by single-run nanoLC-MS/MS using ultralong gradients | Nature Protocols https://www.nature.com/articles/nprot.2012.036 (accessed Mar 31, 2020).
(77) Method of preventing contamination of a chromatography column - Sarasep, Inc. http://www.freepatentsonline.com/5338448.html (accessed Mar 31, 2020).

(78) Dragacci, S.; Grosso, F.; Gilbert, J.; Collaborators:; Agnedal, M.; Hyndrick, L.; Jamet, G.; Jorgensen, K.; Miller, J.; Oliveira Palavras, L.; Pittet, A.; Rousi, V.; Sharron, P.; Sizoo, E. A.; Spott, M.; Strassmeier, E. Immunoaffinity Column Cleanup with Liquid Chromatography for Determination of Aflatoxin M1 in Liquid Milk: Collaborative Study. *J. AOAC Int.* **2001**, *84* (2), 437–443. https://doi.org/10.1093/jaoac/84.2.437.
(79) King, R.; Bonfiglio, R.; Fernandez-Metzler, C.; Miller-Stein, C.; Olah, T. Mechanistic Investigation of Ionization Suppression in Electrospray Ionization. *J. Am. Soc. Mass Spectrom.* **2000**, *11* (11), 942–950. https://doi.org/10.1016/S1044-0305(00)00163-X.
(80) Ho, C.; Lam, C.; Chan, M.; Cheung, R.; Law, L.; Lit, L.; Ng, K.; Suen, M.; Tai, H. Electrospray Ionisation Mass Spectrometry: Principles and Clinical Applications. *Clin. Biochem. Rev.* **2003**, *24* (1), 3–12.
(81) Rohner, T. C.; Lion, N.; Girault, H. H. Electrochemical and Theoretical Aspects of Electrospray Ionisation. *Phys. Chem. Chem. Phys.* **2004**, *6* (12), 3056. https://doi.org/10.1039/b316836k.
(82) Gibbons, B. C.; Chambers, M. C.; Monroe, M. E.; Tabb, D. L.; Payne, S. H. Correcting Systematic Bias and Instrument Measurement Drift with MzRefinery. *Bioinformatics* **2015**, *31* (23), 3838–3840. https://doi.org/10.1093/bioinformatics/btv437.
(83) Tabb, D. L.; Vega-Montoto, L.; Rudnick, P. A.; Variyath, A. M.; Ham, A.-J. L.; Bunk, D. M.; Kilpatrick, L. E.; Billheimer, D. D.; Blackman, R. K.; Cardasis, H. L.; Carr, S. A.; Clauser, K. R.; Jaffe, J. D.; Kowalski, K. A.; Neubert, T. A.; Regnier, F. E.; Schilling, B.; Tegeler, T. J.; Wang, M.; Wang, P.; Whiteaker, J. R.; Zimmerman, L. J.; Fisher, S. J.; Gibson, B. W.; Kinsinger, C. R.; Mesri, M.; Rodriguez, H.; Stein, S. E.; Tempst, P.; Paulovich, A. G.; Liebler, D. C.; Spiegelman, C. Repeatability and Reproducibility in Proteomic Identifications by Liquid Chromatography—Tandem Mass Spectrometry. *J. Proteome Res.* **2010**, *9* (2), 761. https://doi.org/10.1021/pr9006365.
(84) Egertson, J. D.; Kuehn, A.; Merrihew, G. E.; Bateman, N. W.; MacLean, B. X.; Ting, Y. S.; Canterbury, J. D.; Marsh, D. M.; Kellmann, M.; Zabrouskov, V.; Wu, C. C.; MacCoss, M. J. Multiplexed MS/MS for Improved Data-Independent Acquisition. *Nat. Methods* **2013**, *10* (8), 744–746. https://doi.org/10.1038/nmeth.2528.
(85) Amodei, D.; Egertson, J.; MacLean, B. X.; Johnson, R.; Merrihew, G. E.; Keller, A.; Marsh, D.; Vitek, O.; Mallick, P.; MacCoss, M. J. Improving Precursor Selectivity in Data-Independent Acquisition Using Overlapping Windows. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (4), 669–684. https://doi.org/10.1021/jasms.8b05980.
(86) Zhang, Y.; Bilbao, A.; Bruderer, T.; Luban, J.; Strambio-De-Castillia, C.; Lisacek, F.; Hopfgartner, G.; Varesio, E. The Use of Variable Q1 Isolation Windows Improves Selectivity in LC–SWATH–MS Acquisition. *J. Proteome Res.* **2015**, *14* (10), 4359–4371. https://doi.org/10.1021/acs.jproteome.5b00543.
(87) Quantitative Mass Spectrometry Independence from Matrix Effects and Detector Saturation Achieved by Flow Injection Analysis with Real-Time Infinite Dilution | Analytical Chemistry https://pubs.acs.org/doi/full/10.1021/ac402567w?casa_token=bO5xUZ8ih5UAAAAA:EzkBcw07wGtRnb3CQsngZ69Ezhk3vkbySZN1NQMt5XK7zrCwbpFOIKOwDFWB7T0DjqSG3MH5_YBFJYQ (accessed Jun 16, 2020).
(88) Mass Spectrometry Sample Preparation Procedure for Protein Samples - ZA https://www.thermofisher.com/za/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/protein-biology-application-notes/mass-spectrometry-sample-preparation-procedure-protein-samples.html (accessed May 25, 2020).
(89) Doneanu, C.; Yang, H.; Rainville, P.; Bouvier, E.; Plumb, R. Optimization of Trypsin Digestion for MRM Quantification of Therapeutic Proteins in Serum. 7.
(90) Deng, Y.; Butré, C. I.; Wierenga, P. A. Influence of Substrate Concentration on the Extent

of Protein Enzymatic Hydrolysis. *Int. Dairy J.* **2018**, *86*, 39–48. https://doi.org/10.1016/j.idairyj.2018.06.018.

(91) Zhong, J.; Krawczyk, S. A.; Chaerkady, R.; Huang, H.; Goel, R.; Bader, J. S.; Wong, G. W.; Corkey, B. E.; Pandey, A. Temporal Profiling of the Secretome during Adipogenesis in Humans. *J. Proteome Res.* **2010**, *9* (10), 5228–5238. https://doi.org/10.1021/pr100521c.

(92) Hall, S. L.; Hester, S.; Griffin, J. L.; Lilley, K. S.; Jackson, A. P. The Organelle Proteome of the DT40 Lymphocyte Cell Line. *Mol. Cell. Proteomics MCP* **2009**, *8* (6), 1295–1305. https://doi.org/10.1074/mcp.M800394-MCP200.

(93) Christoforou, A. L.; Lilley, K. S. Isobaric Tagging Approaches in Quantitative Proteomics: The Ups and Downs. *Anal. Bioanal. Chem.* **2012**, *404* (4), 1029–1037. https://doi.org/10.1007/s00216-012-6012-9.

(94) Bittremieux, W.; Tabb, D. L.; Impens, F.; Staes, A.; Timmerman, E.; Martens, L.; Laukens, K. Quality Control in Mass Spectrometry-Based Proteomics. *Mass Spectrom. Rev.* **2017**. https://doi.org/10.1002/mas.21544.

(95) Ramus, C.; Hovasse, A.; Marcellin, M.; Hesse, A.-M.; Mouton-Barbosa, E.; Bouyssié, D.; Vaca, S.; Carapito, C.; Chaoui, K.; Bruley, C.; Garin, J.; Cianférani, S.; Ferro, M.; Van Dorssaeler, A.; Burlet-Schiltz, O.; Schaeffer, C.; Couté, Y.; Gonzalez de Peredo, A. Benchmarking Quantitative Label-Free LC–MS Data Processing Workflows Using a Complex Spiked Proteomic Standard Dataset. *J. Proteomics* **2016**, *132*, 51–62. https://doi.org/10.1016/j.jprot.2015.11.011.

(96) Ramus, C.; Hovasse, A.; Marcellin, M.; Hesse, A.-M.; Mouton-Barbosa, E.; Bouyssié, D.; Vaca, S.; Carapito, C.; Chaoui, K.; Bruley, C.; Garin, J.; Cianférani, S.; Ferro, M.; Dorssaeler, A. V.; Burlet-Schiltz, O.; Schaeffer, C.; Couté, Y.; Gonzalez de Peredo, A. Spiked Proteomic Standard Dataset for Testing Label-Free Quantitative Software and Statistical Methods. *Data Brief* **2016**, *6*, 286–294. https://doi.org/10.1016/j.dib.2015.11.063.

(97) Escher, C.; Reiter, L.; MacLean, B.; Ossola, R.; Herzog, F.; Chilton, J.; MacCoss, M. J.; Rinner, O. Using IRT, a Normalized Retention Time for More Targeted Measurement of Peptides. *Proteomics* **2012**, *12* (8), 1111–1121. https://doi.org/10.1002/pmic.201100463.

(98) Biognosys - Next generation proteomics https://biognosys.com/shop/quic (accessed May 25, 2020).

(99) Ma, Z.-Q.; Polzin, K. O.; Dasari, S.; Chambers, M. C.; Schilling, B.; Gibson, B. W.; Tran, B. Q.; Vega-Montoto, L.; Liebler, D. C.; Tabb, D. L. QuaMeter: Multivendor Performance Metrics for LC–MS/MS Proteomics Instrumentation. *Anal. Chem.* **2012**, *84* (14), 5845–5850. https://doi.org/10.1021/ac300629p.

(100) Wang, X.; Chambers, M. C.; Vega-Montoto, L. J.; Bunk, D. M.; Stein, S. E.; Tabb, D. L. QC Metrics from CPTAC Raw LC-MS/MS Data Interpreted through Multivariate Statistics. *Anal. Chem.* **2014**, *86* (5), 2497–2509. https://doi.org/10.1021/ac4034455.

(101) Bielow, C.; Mastrobuoni, G.; Kempa, S. Proteomics Quality Control: Quality Control Software for MaxQuant Results. *J. Proteome Res.* **2016**, *15* (3), 777–787. https://doi.org/10.1021/acs.jproteome.5b00780.

(102) Scheltema, R. A.; Mann, M. SprayQc: A Real-Time LC-MS/MS Quality Monitoring System to Maximize Uptime Using off the Shelf Components. *J. Proteome Res.* **2012**, *11* (6), 3458–3466. https://doi.org/10.1021/pr201219e.

(103) MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J. Skyline: An Open Source Document Editor for Creating and Analyzing Targeted Proteomics Experiments. *Bioinformatics* **2010**, *26* (7), 966–968. https://doi.org/10.1093/bioinformatics/btq054.

(104) Metriculator: quality assessment for mass spectrometry-based proteomics. - PubMed - NCBI https://www.ncbi.nlm.nih.gov/pubmed/24002108 (accessed Jul 31, 2019).

(105) Pfeuffer, J.; Sachsenberg, T.; Alka, O.; Walzer, M.; Fillbrunn, A.; Nilse, L.; Schilling, O.; Reinert, K.; Kohlbacher, O. OpenMS – A Platform for Reproducible Analysis of Mass

Spectrometry Data. *J. Biotechnol.* **2017**, *261*, 142–148. https://doi.org/10.1016/j.jbiotec.2017.05.016.

(106) Bereman, M. S.; Johnson, R.; Bollinger, J.; Boss, Y.; Shulman, N.; MacLean, B.; Hoofnagle, A. N.; MacCoss, M. J. Implementation of Statistical Process Control for Proteomic Experiments via LC MS/MS. *J. Am. Soc. Mass Spectrom.* **2014**, *25* (4), 581–587. https://doi.org/10.1007/s13361-013-0824-5.

(107) Huffman, G.; Specht, H.; Chen, A. T.; Slavov, N. DO-MS: Data-Driven Optimization of Mass Spectrometry Methods. *bioRxiv* **2019**. https://doi.org/10.1101/512152.

(108) Dogu, E.; Taheri, S. M.; Olivella, R.; Marty, F.; Lienert, I.; Reiter, L.; Sabido, E.; Vitek, O. MSstatsQC 2.0: R/Bioconductor Package for Statistical Quality Control of Mass Spectrometry-Based Proteomics Experiments. *J. Proteome Res.* **2019**, *18* (2), 678–686. https://doi.org/10.1021/acs.jproteome.8b00732.

(109) Bittremieux, W.; Willems, H.; Kelchtermans, P.; Martens, L.; Laukens, K.; Valkenborg, D. IMonDB: Mass Spectrometry Quality Control through Instrument Monitoring. *J. Proteome Res.* **2015**, *14* (5), 2360–2366. https://doi.org/10.1021/acs.jproteome.5b00127.

(110) Trachsel, C.; Panse, C.; Kockmann, T.; Wolski, W. E.; Grossmann, J.; Schlapbach, R. RawDiag: An R Package Supporting Rational LC-MS Method Optimization for Bottom-up Proteomics. *J. Proteome Res.* **2018**, *17* (8), 2908–2914. https://doi.org/10.1021/acs.jproteome.8b00173.

(111) Pichler, P.; Mazanek, M.; Dusberger, F.; Weilnböck, L.; Huber, C. G.; Stingl, C.; Luider, T. M.; Straube, W. L.; Köcher, T.; Mechtler, K. SIMPATIQCO: A Server-Based Software Suite Which Facilitates Monitoring the Time Course of LC–MS Performance Metrics on Orbitrap Instruments. *J. Proteome Res.* **2012**, *11* (11), 5540–5547. https://doi.org/10.1021/pr300163u.

(112) Chiva, C.; Olivella, R.; Borràs, E.; Espadas, G.; Pastor, O.; Solé, A.; Sabidó, E. QCloud: A Cloud-Based Quality Control System for Mass Spectrometry-Based Proteomics Laboratories. *PLoS ONE* **2018**, *13* (1). https://doi.org/10.1371/journal.pone.0189209.

(113) Stratton, K. G.; Webb-Robertson, B.-J. M.; McCue, L. A.; Stanfill, B.; Claborne, D.; Godinez, I.; Johansen, T.; Thompson, A. M.; Burnum-Johnson, K. E.; Waters, K. M.; Bramer, L. M. PmartR: Quality Control and Statistics for Mass Spectrometry-Based Biological Data. *J. Proteome Res.* **2019**, *18* (3), 1418–1425. https://doi.org/10.1021/acs.jproteome.8b00760.

(114) Luan, H.; Ji, F.; Chen, Y.; Cai, Z. StatTarget: A Streamlined Tool for Signal Drift Correction and Interpretations of Quantitative Mass Spectrometry-Based Omics Data. *Anal. Chim. Acta* **2018**, *1036*, 66–72. https://doi.org/10.1016/j.aca.2018.08.002.

(115) Kim, T.; Chen, I. R.; Parker, B. L.; Humphrey, S. J.; Crossett, B.; Cordwell, S. J.; Yang, P.; Yang, J. Y. H. QCMAP: An Interactive Web- Tool for Performance Diagnosis and Prediction of LC- MS Systems https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.201900068 (accessed Jul 31, 2019). https://doi.org/10.1002/pmic.201900068.

(116) Allen, C.; Mehler, D. M. A. Open Science Challenges, Benefits and Tips in Early Career and Beyond. *PLOS Biol.* **2019**, *17* (5), e3000246. https://doi.org/10.1371/journal.pbio.3000246.

(117) Deutsch, E. W.; Csordas, A.; Sun, Z.; Jarnuczak, A.; Perez-Riverol, Y.; Ternent, T.; Campbell, D. S.; Bernal-Llinares, M.; Okuda, S.; Kawano, S.; Moritz, R. L.; Carver, J. J.; Wang, M.; Ishihama, Y.; Bandeira, N.; Hermjakob, H.; Vizcaíno, J. A. The ProteomeXchange Consortium in 2017: Supporting the Cultural Change in Proteomics Public Data Deposition. *Nucleic Acids Res.* **2017**, *45* (D1), D1100–D1106. https://doi.org/10.1093/nar/gkw936.

(118) Molloy, J. C. The Open Knowledge Foundation: Open Data Means Better Science. *PLOS Biol.* **2011**, *9* (12), e1001195. https://doi.org/10.1371/journal.pbio.1001195.

(119) Resnik, D. B.; Morales, M.; Landrum, R.; Shi, M.; Minnier, J.; Vasilevsky, N. A.; Champieux, R. E. Effect of Impact Factor and Discipline on Journal Data Sharing Policies. *Account. Res.* **2019**, *26* (3), 139–156. https://doi.org/10.1080/08989621.2019.1591277.

(120) Moriya, Y.; Kawano, S.; Okuda, S.; Watanabe, Y.; Matsumoto, M.; Takami, T.; Kobayashi, D.; Yamanouchi, Y.; Araki, N.; Yoshizawa, A. C.; Tabata, T.; Iwasaki, M.; Sugiyama, N.; Tanaka, S.; Goto, S.; Ishihama, Y. The JPOST Environment: An Integrated Proteomics Data Repository and Database. *Nucleic Acids Res.* **2019**, *47* (Database issue), D1218–D1224. https://doi.org/10.1093/nar/gky899.

(121) Jones, P.; Côté, R. The PRIDE Proteomics Identifications Database: Data Submission, Query, and Dataset Comparison. *Methods Mol. Biol. Clifton NJ* **2008**, *484*, 287–303. https://doi.org/10.1007/978-1-59745-398-1_19.

(122) Deutsch, E. W.; Lam, H.; Aebersold, R. PeptideAtlas: A Resource for Target Selection for Emerging Targeted Proteomics Workflows. *EMBO Rep.* **2008**, *9* (5), 429–434. https://doi.org/10.1038/embor.2008.56.

(123) Doerr, A. Proteomics Data Reuse with MassIVE-KB. *Nat. Methods* **2019**, *16* (1), 26–26. https://doi.org/10.1038/s41592-018-0283-9.

(124) Farrah, T.; Deutsch, E. W.; Kreisberg, R.; Sun, Z.; Campbell, D. S.; Mendoza, L.; Kusebauch, U.; Brusniak, M.-Y.; Hüttenhain, R.; Schiess, R.; Selevsek, N.; Aebersold, R.; Moritz, R. L. PASSEL: The PeptideAtlas SRM Experiment Library. *Proteomics* **2012**, *12* (8). https://doi.org/10.1002/pmic.201100515.

(125) Sharma, V.; Eckels, J.; Schilling, B.; Ludwig, C.; Jaffe, J. D.; MacCoss, M. J.; MacLean, B. Panorama Public: A Public Repository for Quantitative Data Sets Processed in Skyline. *Mol. Cell. Proteomics MCP* **2018**, *17* (6), 1239–1244. https://doi.org/10.1074/mcp.RA117.000543.

(126) Ma, J.; Chen, T.; Wu, S.; Yang, C.; Bai, M.; Shu, K.; Li, K.; Zhang, G.; Jin, Z.; He, F.; Hermjakob, H.; Zhu, Y. IProX: An Integrated Proteome Resource. *Nucleic Acids Res.* **2019**, *47* (Database issue), D1211–D1217. https://doi.org/10.1093/nar/gky869.

(127) Mehaffy, M. C.; Kruh-Garcia, N. A.; Dobos, K. M. Prospective on Mycobacterium Tuberculosis Proteomics. *J. Proteome Res.* **2012**, *11* (1), 17–25. https://doi.org/10.1021/pr2008658.

(128) Schubert, O. T.; Mouritsen, J.; Ludwig, C.; Röst, H. L.; Rosenberger, G.; Arthur, P. K.; Claassen, M.; Campbell, D. S.; Sun, Z.; Farrah, T.; Gengenbacher, M.; Maiolica, A.; Kaufmann, S. H. E.; Moritz, R. L.; Aebersold, R. The Mtb Proteome Library: A Resource of Assays to Quantify the Complete Proteome of Mycobacterium Tuberculosis. *Cell Host Microbe* **2013**, *13* (5), 602–612. https://doi.org/10.1016/j.chom.2013.04.008.

(129) Veglia, F.; Perego, M.; Gabrilovich, D. Myeloid-Derived Suppressor Cells Coming of Age. *Nat. Immunol.* **2018**, *19* (2), 108–119. https://doi.org/10.1038/s41590-017-0022-x.

(130) Gabrilovich, D. I.; Nagaraj, S. Myeloid-Derived Suppressor Cells as Regulators of the Immune System. *Nat. Rev. Immunol.* **2009**, *9* (3), 162–174. https://doi.org/10.1038/nri2506.

(131) Jayashankar, L.; Hafner, R. Adjunct Strategies for Tuberculosis Vaccines: Modulating Key Immune Cell Regulatory Mechanisms to Potentiate Vaccination. *Front. Immunol.* **2016**, *7*, 577. https://doi.org/10.3389/fimmu.2016.00577.

(132) du Plessis, N.; Kotze, L. A.; Leukes, V.; Walzl, G. Translational Potential of Therapeutics Targeting Regulatory Myeloid Cells in Tuberculosis. *Front. Cell. Infect. Microbiol.* **2018**, *8*, 332. https://doi.org/10.3389/fcimb.2018.00332.

(133) Bruger, A.; Dorhoi, A.; Esendagli, G.; Barczyk-Kahlert, K.; van der Bruggen, P.; Lipoldova, M.; Perecko, T.; Santibanez, J.; Saraiva, M.; Van Ginderachter, J.; Brandau, S. How to Measure the Immunosuppressive Activity of MDSC: Assays, Problems and Potential Solutions. *Cancer Immunol. Immunother.* **2019**, *68* (4), 631–644. https://doi.org/10.1007/s00262-018-2170-8.

(134) Gato, M.; Blanco-Luquin, I.; Zudaire, M.; de Morentin, X. M.; Perez-Valderrama, E.; Zabaleta, A.; Kochan, G.; Escors, D.; Fernandez-Irigoyen, J.; Santamaría, E. Drafting the Proteome Landscape of Myeloid-Derived Suppressor Cells. *PROTEOMICS* **2016**, *16* (2), 367–378. https://doi.org/10.1002/pmic.201500229.

(135) Cassetta, L.; Baekkevold, E. S.; Brandau, S.; Bujko, A.; Cassatella, M. A.; Dorhoi, A.; Krieg, C.; Lin, A.; Loré, K.; Marini, O.; Pollard, J. W.; Roussel, M.; Scapini, P.; Umansky, V.; Adema, G. J. Deciphering Myeloid-Derived Suppressor Cells: Isolation and Markers in Humans, Mice and Non-Human Primates. *Cancer Immunol. Immunother.* **2019**, *68* (4), 687–697. https://doi.org/10.1007/s00262-019-02302-2.

(136) Burke, M.; Choksawangkarn, W.; Edwards, N.; Ostrand-Rosenberg, S.; Fenselau, C. Exosomes from Myeloid-Derived Suppressor Cells Carry Biologically Active Proteins. *J. Proteome Res.* **2014**, *13* (2), 836–843. https://doi.org/10.1021/pr400879c.

(137) Geis-Asteggiante, L.; Belew, A. T.; Clements, V. K.; Edwards, N. J.; Ostrand-Rosenberg, S.; El-Sayed, N. M.; Fenselau, C. Differential Content of Proteins, MRNAs, and MiRNAs Suggests That MDSC and Their Exosomes May Mediate Distinct Immune Suppressive Functions. *J. Proteome Res.* **2018**, *17* (1), 486–498. https://doi.org/10.1021/acs.jproteome.7b00646.

(138) Xiang, X.; Poliakov, A.; Liu, C.; Liu, Y.; Deng, Z.; Cheng, Z.; Shah, S.; Wang, G.; Zhang, L.; Grizzle, W.; Mobley, J.; Zhang, H. Induction of Myeloid-Derived Suppressor Cells by Tumor Exosomes. *Int. J. Cancer* **2009**, *124* (11), 2621–2633. https://doi.org/10.1002/ijc.24249.

(139) Beaudry, F.; Vachon, P. Determination of Substance P in Rat Spinal Cord by High-Performance Liquid Chromatography Electrospray Quadrupole Ion Trap Mass Spectrometry. *Biomed. Chromatogr.* **2006**, *20* (12), 1344–1350. https://doi.org/10.1002/bmc.703.

(140) Mallet, C. R.; Lu, Z.; Mazzeo, J. R. A Study of Ion Suppression Effects in Electrospray Ionization from Mobile Phase Additives and Solid-Phase Extracts. *Rapid Commun. Mass Spectrom.* **2004**, *18* (1), 49–58. https://doi.org/10.1002/rcm.1276.

(141) Kelkar, D. S.; Kumar, D.; Kumar, P.; Balakrishnan, L.; Muthusamy, B.; Yadav, A. K.; Shrivastava, P.; Marimuthu, A.; Anand, S.; Sundaram, H.; Kingsbury, R.; Harsha, H. C.; Nair, B.; Prasad, T. S.; Chauhan, D. S.; Katoch, K.; Katoch, V. M.; Kumar, P.; Chaerkady, R.; Ramachandran, S.; Dash, D.; Pandey, A. Proteogenomic Analysis of Mycobacterium Tuberculosis by High Resolution Mass Spectrometry. *Mol. Cell. Proteomics MCP* **2011**, *10* (12), M111.011627-M111.011627. https://doi.org/10.1074/mcp.M111.011627.

(142) Heunis, T.; Dippenaar, A.; Warren, R. M.; van Helden, P. D.; van der Merwe, R. G.; Gey van Pittius, N. C.; Pain, A.; Sampson, S. L.; Tabb, D. L. Proteogenomic Investigation of Strain Variation in Clinical *Mycobacterium Tuberculosis* Isolates. *J. Proteome Res.* **2017**, *16* (10), 3841–3851. https://doi.org/10.1021/acs.jproteome.7b00483.

(143) Liechtenstein, T.; Perez-Janices, N.; Gato, M.; Caliendo, F.; Kochan, G.; Blanco-Luquin, I.; Arce, F.; Guerrero-Setas, D.; Fernandez-Irigoyen, J.; Santamaria, E.; Breckpot, K.; Escors, D. A Highly Efficient Tumor-Infiltrating MDSC Differentiation System for Discovery of Anti-Neoplastic Targets, Which Circumvents the Need for Tumor Establishment in Mice. *Oncotarget* **2014**, *5* (17). https://doi.org/10.18632/oncotarget.2279.

(144) Chornoguz, O.; Grmai, L.; Sinha, P.; Artemenko, K. A.; Zubarev, R. A.; Ostrand-Rosenberg, S. Proteomic Pathway Analysis Reveals Inflammation Increases Myeloid-Derived Suppressor Cell Resistance to Apoptosis. *Mol. Cell. Proteomics* **2011**, *10* (3), M110.002980. https://doi.org/10.1074/mcp.M110.002980.

(145) Zöller, M.; Zhao, K.; Kutlu, N.; Bauer, N.; Provaznik, J.; Hackert, T.; Schnölzer, M. Immunoregulatory Effects of Myeloid-Derived Suppressor Cell Exosomes in Mouse Model of Autoimmune Alopecia Areata. *Front. Immunol.* **2018**, *9*. https://doi.org/10.3389/fimmu.2018.01279.

(146) Wojtuszkiewicz, A.; Schuurhuis, G. J.; Kessler, F. L.; Piersma, S. R.; Knol, J. C.; Pham, T. V.; Jansen, G.; Musters, R. J. P.; van Meerloo, J.; Assaraf, Y. G.; Kaspers, G. J. L.; Zweegman, S.; Cloos, J.; Jimenez, C. R. Exosomes Secreted by Apoptosis-Resistant Acute Myeloid Leukemia (AML) Blasts Harbor Regulatory Network Proteins Potentially Involved in Antagonism of Apoptosis. *Mol. Cell. Proteomics* **2016**, *15* (4), 1281–1298. https://doi.org/10.1074/mcp.M115.052944.

(147) Zhang, J.; Lu, S.; Zhou, Y.; Meng, K.; Chen, Z.; Cui, Y.; Shi, Y.; Wang, T.; He, Q.-Y. Motile Hepatocellular Carcinoma Cells Preferentially Secret Sugar Metabolism Regulatory Proteins via Exosomes. *PROTEOMICS* **2017**, *17* (13–14), 1700103. https://doi.org/10.1002/pmic.201700103.

(148) Zhao, X.; Wu, Y.; Duan, J.; Ma, Y.; Shen, Z.; Wei, L.; Cui, X.; Zhang, J.; Xie, Y.; Liu, J. Quantitative Proteomic Analysis of Exosome Protein Content Changes Induced by Hepatitis B Virus in Huh-7 Cells Using SILAC Labeling and LC–MS/MS. *J. Proteome Res.* **2014**, *13* (12), 5391–5402. https://doi.org/10.1021/pr5008703.

(149) Diaz, G.; Wolfe, L. M.; Kruh-Garcia, N. A.; Dobos, K. M. Changes in the Membrane-Associated Proteins of Exosomes Released from Human Macrophages after Mycobacterium Tuberculosis Infection. *Sci. Rep.* **2016**, *6* (1). https://doi.org/10.1038/srep37975.

(150) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Römpp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.; Souda, P.; Hermjakob, H.; Binz, P.-A.; Deutsch, E. W. MzML—a Community Standard for Mass Spectrometry Data. *Mol. Cell. Proteomics MCP* **2011**, *10* (1). https://doi.org/10.1074/mcp.R110.000133.

(151) Holman, J. D.; Tabb, D. L.; Mallick, P. Employing ProteoWizard to Convert Raw Mass Spectrometry Data: Employing ProteoWizard to Convert Raw Mass Spectrometry Data. In *Current Protocols in Bioinformatics*; Bateman, A., Pearson, W. R., Stein, L. D., Stormo, G. D., Yates, J. R., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2014; p 13.24.1-13.24.9. https://doi.org/10.1002/0471250953.bi1324s46.

(152) Modern Applied Statistics with S, 4th ed http://www.stats.ox.ac.uk/pub/MASS4/ (accessed Oct 7, 2019).

(153) The Personality Project's Guide to R https://personality-project.org/r/psych/ (accessed Jul 7, 2020).

(154) Sarkar, D. *Lattice: Multivariate Data Visualization with R*; Use R!; Springer-Verlag: New York, 2008. https://doi.org/10.1007/978-0-387-75969-2.

(155) Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using Lme4. *J. Stat. Softw.* **2015**, *67* (1), 1–48. https://doi.org/10.18637/jss.v067.i01.

(156) Fox, J.; Weisberg, S.; Price, B.; Adler, D.; Bates, D.; Baud-Bovy, G.; Bolker, B.; Ellison, S.; Firth, D.; Friendly, M.; Gorjanc, G.; Graves, S.; Heiberger, R.; Krivitsky, P.; Laboissiere, R.; Maechler, M.; Monette, G.; Murdoch, D.; Nilsson, H.; Ogle, D.; Ripley, B.; Venables, W.; Walker, S.; Winsemius, D.; Zeileis, A.; R-Core. *Car: Companion to Applied Regression*; 2020.

(157) Galili, T. Dendextend: An R Package for Visualizing, Adjusting and Comparing Trees of Hierarchical Clustering. *Bioinformatics* **2015**, *31* (22), 3718–3720. https://doi.org/10.1093/bioinformatics/btv428.

(158) Tidyverse https://www.tidyverse.org/packages/#installation-and-use (accessed Jul 7, 2020).

(159) Tang, Y.; Horikoshi, M.; Li, W. Ggfortify: Unified Interface to Visualize Statistical Results of Popular R Packages. *R J.* **2016**, *8*, 478–489. https://doi.org/10.32614/RJ-2016-060.

(160) Nested anova - Handbook of Biological Statistics http://www.biostathandbook.com/nestedanova.html (accessed Feb 6, 2020).

(161) Multiple comparisons - Handbook of Biological Statistics

http://www.biostathandbook.com/multiplecomparisons.html (accessed Feb 10, 2020).
(162) Heinze, G.; Wallisch, C.; Dunkler, D. Variable Selection – A Review and Recommendations for the Practicing Statistician. *Biom. J. Biom. Z.* **2018**, *60* (3), 431–449. https://doi.org/10.1002/bimj.201700067.
(163) Ferré, L. Selection of Components in Principal Component Analysis: A Comparison of Methods. *Comput. Stat. Data Anal.* **1995**, *19* (6), 669–682. https://doi.org/10.1016/0167-9473(94)00020-J.
(164) Seo, S. A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets http://d-scholarship.pitt.edu/7948/ (accessed Jan 11, 2020).
(165) Estivill-Castro, V.; Yang, J. Fast and Robust General Purpose Clustering Algorithms. *Data Min. Knowl. Discov.* **2004**, *8* (2), 127–150. https://doi.org/10.1023/B:DAMI.0000015869.08323.b3.
(166) Kim, S.; Pevzner, P. A. Universal Database Search Tool for Proteomics. *Nat. Commun.* **2014**, *5*, 5277. https://doi.org/10.1038/ncomms6277.
(167) Jones, A. R.; Eisenacher, M.; Mayer, G.; Kohlbacher, O.; Siepen, J.; Hubbard, S. J.; Selley, J. N.; Searle, B. C.; Shofstahl, J.; Seymour, S. L.; Julian, R.; Binz, P.-A.; Deutsch, E. W.; Hermjakob, H.; Reisinger, F.; Griss, J.; Vizcaíno, J. A.; Chambers, M.; Pizarro, A.; Creasy, D. The MzIdentML Data Standard for Mass Spectrometry-Based Proteomics Results. *Mol. Cell. Proteomics MCP* **2012**, *11* (7), M111.014381. https://doi.org/10.1074/mcp.M111.014381.
(168) Amidan, B. G.; Orton, D. J.; LaMarche, B. L.; Monroe, M. E.; Moore, R. J.; Venzin, A. M.; Smith, R. D.; Sego, L. H.; Tardiff, M. F.; Payne, S. H. Signatures for Mass Spectrometry Data Quality. *J. Proteome Res.* **2014**, *13* (4), 2215–2222. https://doi.org/10.1021/pr401143e.
(169) Solovyeva, E. M.; Lobas, A. A.; Kopylov, A. T.; Gorshkov, M. V. Semi-Supervised Quality Control Method for Proteome Analyses Based on Tandem Mass Spectrometry. *Int. J. Mass Spectrom.* **2018**, *427*, 59–64. https://doi.org/10.1016/j.ijms.2017.09.008.
(170) Bittremieux, W.; Meysman, P.; Martens, L.; Valkenborg, D.; Laukens, K. Unsupervised Quality Assessment of Mass Spectrometry Proteomics Experiments by Multivariate Quality Control Metrics. *J. Proteome Res.* **2016**, *15* (4), 1300–1307. https://doi.org/10.1021/acs.jproteome.6b00028.
(171) Wang, X.; Chambers, M. C.; Vega-Montoto, L. J.; Bunk, D. M.; Stein, S. E.; Tabb, D. L. QC Metrics from CPTAC Raw LC-MS/MS Data Interpreted through Multivariate Statistics. *Anal. Chem.* **2014**, *86* (5), 2497–2509. https://doi.org/10.1021/ac4034455.
(172) Manadas, B.; Mendes, V. M.; English, J.; Dunn, M. J. Peptide Fractionation in Proteomics Approaches. *Expert Rev. Proteomics* **2010**, *7* (5), 655–663. https://doi.org/10.1586/epr.10.46.
(173) Mostovenko, E.; Hassan, C.; Rattke, J.; Deelder, A. M.; van Veelen, P. A.; Palmblad, M. Comparison of Peptide and Protein Fractionation Methods in Proteomics. *EuPA Open Proteomics* **2013**, *1*, 30–37. https://doi.org/10.1016/j.euprot.2013.09.001.
(174) Comparison of protein and peptide prefractionation methods for the shotgun proteomic analysis of Synechocystis sp. PCC 6803 - Gan - 2005 - PROTEOMICS - Wiley Online Library https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.200401266 (accessed Aug 13, 2019).
(175) Kelkar, D. S.; Kumar, D.; Kumar, P.; Balakrishnan, L.; Muthusamy, B.; Yadav, A. K.; Shrivastava, P.; Marimuthu, A.; Anand, S.; Sundaram, H.; Kingsbury, R.; Harsha, H. C.; Nair, B.; Prasad, T. S. K.; Chauhan, D. S.; Katoch, K.; Katoch, V. M.; Kumar, P.; Chaerkady, R.; Ramachandran, S.; Dash, D.; Pandey, A. Proteogenomic Analysis of *Mycobacterium Tuberculosis* By High Resolution Mass Spectrometry. *Mol. Cell. Proteomics* **2011**, *10* (12), M111.011627. https://doi.org/10.1074/mcp.M111.011627.
(176) Shlens, J. A Tutorial on Principal Component Analysis. *ArXiv14041100 Cs Stat* **2014**.

(177) Zhao, X.; Wu, Y.; Duan, J.; Ma, Y.; Shen, Z.; Wei, L.; Cui, X.; Zhang, J.; Xie, Y.; Liu, J. Quantitative Proteomic Analysis of Exosome Protein Content Changes Induced by Hepatitis B Virus in Huh-7 Cells Using SILAC Labeling and LC–MS/MS. *J. Proteome Res.* **2014**, *13* (12), 5391–5402. https://doi.org/10.1021/pr5008703.
(178) Fisher, R. A. Design of Experiments. *Br. Med. J.* **1936**, *1* (3923), 554.
(179) Albar, J.-P.; Canals, F. Standardization and Quality Control in Proteomics. *J. Proteomics* **2013**, *95*, 1–2. https://doi.org/10.1016/j.jprot.2013.11.002.
(180) Bramwell, D. An Introduction to Statistical Process Control in Research Proteomics. *J. Proteomics* **2013**, *95*, 3–21. https://doi.org/10.1016/j.jprot.2013.06.010.
(181) Quality meets quantity – quality control, data standards and repositories https://onlinelibrary-wiley-com.ez.sun.ac.za/doi/epdf/10.1002/pmic.201000441 (accessed Mar 31, 2020).
(182) Schubert, O. T.; Ludwig, C.; Kogadeeva, M.; Zimmermann, M.; Rosenberger, G.; Gengenbacher, M.; Gillet, L. C.; Collins, B. C.; Röst, H. L.; Kaufmann, S. H. E.; Sauer, U.; Aebersold, R. Absolute Proteome Composition and Dynamics during Dormancy and Resuscitation of Mycobacterium Tuberculosis. *Cell Host Microbe* **2015**, *18* (1), 96–108. https://doi.org/10.1016/j.chom.2015.06.001.
(183) Kim, Y. J.; Sweet, S. M. M.; Egertson, J. D.; Sedgewick, A. J.; Woo, S.; Liao, W.; Merrihew, G. E.; Searle, B. C.; Vaske, C.; Heaton, R.; MacCoss, M. J.; Hembrough, T. Data-Independent Acquisition Mass Spectrometry To Quantify Protein Levels in FFPE Tumor Biopsies for Molecular Diagnostics. *J. Proteome Res.* **2019**, *18* (1), 426–435. https://doi.org/10.1021/acs.jproteome.8b00699.
(184) Prado, R.; Bailão, A.; Silva, L.; de Oliveira, C.; Marques, M.; Silva, L.; Silveira-Lacerda, E.; Lima, A.; Soares, C.; Pereira, M. Proteomic Profile Response of Paracoccidioides Lutzii to the Antifungal Argentilactone. - Abstract - Europe PMC. *Front. Microbiol.* **2015**, No. 6.
(185) Holman, J. D.; Tabb, D. L.; Mallick, P. Employing ProteoWizard to Convert Raw Mass Spectrometry Data: Employing ProteoWizard to Convert Raw Mass Spectrometry Data. In *Current Protocols in Bioinformatics*; Bateman, A., Pearson, W. R., Stein, L. D., Stormo, G. D., Yates, J. R., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2014; p 13.24.1-13.24.9. https://doi.org/10.1002/0471250953.bi1324s46.
(186) HUPO-PSI/mzQC https://github.com/HUPO-PSI/mzQC (accessed Apr 30, 2020).
(187) Seo, S. A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets; 2006.
(188) Perez-Riverol, Y.; Zorin, A.; Dass, G.; Glont, M.; Vizcaíno, J. A.; Jarnuczak, A. F.; Petryszak, R.; Ping, P.; Hermjakob, H. Quantifying the Impact of Public Omics Data. *bioRxiv* **2018**, 282517. https://doi.org/10.1101/282517.
(189) Searle, B. C.; Swearingen, K. E.; Barnes, C. A.; Schmidt, T.; Gessulat, S.; Küster, B.; Wilhelm, M. Generating High Quality Libraries for DIA MS with Empirically Corrected Peptide Predictions. *Nat. Commun.* **2020**, *11* (1), 1548. https://doi.org/10.1038/s41467-020-15346-1.
(190) Rose, R. J.; Damoc, E.; Denisov, E.; Makarov, A.; Heck, A. J. R. High-sensitivty Orbitrap mass analysis of intact macromolecular assemblies https://link-galegroup-com.ez.sun.ac.za/apps/doc/A309791829/AONE?sid=lms (accessed Jan 20, 2020). https://doi.org/10.1038/NMETH.2208.
(191) Biognosys - Next generation proteomics https://biognosys.com/spectronaut-14 (accessed Aug 2, 2020).
(192) Weng, N.; Jian, W. *Targeted Biomarker Quantitation by LC-MS*; John Wiley & Sons, 2017.
(193) Tyanova, S.; Temu, T.; Cox, J. The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics. *Nat. Protoc.* **2016**, *11* (12), 2301-.
(194) Pedrioli, P. G. A. Trans-Proteomic Pipeline: A Pipeline for Proteomic Analysis. In

Proteome Bioinformatics; Hubbard, S. J., Jones, A. R., Eds.; Methods in Molecular Biology™; Humana Press: Totowa, NJ, 2010; pp 213–238. https://doi.org/10.1007/978-1-60761-444-9_15.

(195) Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry | Nature Communications https://www.nature.com/articles/s41467-018-07454-w (accessed May 28, 2020).

(196) Tukey, J., W. *Exploratory Data Analysis*; Addison-Wesley, Reading, MA, 1977.

(197) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(198) *H2oai/H2o-3*; H2O.ai, 2020.

(199) Jones, A. R.; Eisenacher, M.; Mayer, G.; Kohlbacher, O.; Siepen, J.; Shofstahl, J.; Seymour, S. L.; Julian, R.; Binz, P.-A.; Deutsch, E. W.; Hermjakob, H.; Reisinger, F.; Griss, J.; Vizcaíno, J. A.; Chambers, M.; Pizarro, A.; Creasy, D. The MzIdentML Data Standard for Mass Spectrometry-Based Proteomics Results□S. **2012**, 10.

(200) Foster, J. M.; Degroeve, S.; Gatto, L.; Visser, M.; Wang, R.; Griss, J.; Apweiler, R.; Martens, L. A Posteriori Quality Control for the Curation and Reuse of Public Proteomics Data. *PROTEOMICS* **2011**, *11* (11), 2182–2194. https://doi.org/10.1002/pmic.201000602.

(201) Munafò, M. R.; Nosek, B. A.; Bishop, D. V. M.; Button, K. S.; Chambers, C. D.; Percie du Sert, N.; Simonsohn, U.; Wagenmakers, E.-J.; Ware, J. J.; Ioannidis, J. P. A. A Manifesto for Reproducible Science. *Nat. Hum. Behav.* **2017**, *1* (1), 1–9. https://doi.org/10.1038/s41562-016-0021.

(202) Ezkurdia, I.; Vázquez, J.; Valencia, A.; Tress, M. Analyzing the First Drafts of the Human Proteome. *J. Proteome Res.* **2014**, *13* (8), 3854–3855. https://doi.org/10.1021/pr500572z.

(203) Fernández-Costa, C.; Martínez-Bartolomé, S.; McClatchy, D.; Yates, J. R. Improving Proteomics Data Reproducibility with a Dual-Search Strategy. *Anal. Chem.* **2020**, *92* (2), 1697–1701. https://doi.org/10.1021/acs.analchem.9b04955.

(204) Bogdanow, B.; Zauber, H.; Selbach, M. Systematic Errors in Peptide and Protein Identification and Quantification by Modified Peptides. *Mol. Cell. Proteomics MCP* **2016**, *15* (8), 2791–2801. https://doi.org/10.1074/mcp.M115.055103.

(205) Petyuk, V. A.; Gatto, L.; Payne, S. H. Reproducibility and Transparency by Design. *Mol. Cell. Proteomics* **2019**, *18* (8 suppl 1), S202–S204. https://doi.org/10.1074/mcp.IP119.001567.

(206) Barkovits, K.; Pacharra, S.; Pfeiffer, K.; Steinbach, S.; Eisenacher, M.; Marcus, K.; Uszkoreit, J. Reproducibility, Specificity and Accuracy of Relative Quantification Using Spectral Library-Based Data-Independent Acquisition. *Mol. Cell. Proteomics MCP* **2020**, *19* (1), 181–197. https://doi.org/10.1074/mcp.RA119.001714.

(207) Barnouin, K. Guidelines for Experimental Design and Data Analysis of Proteomic Mass Spectrometry-Based Experiments. *Amino Acids* **2011**, *40* (2), 259–260. https://doi.org/10.1007/s00726-010-0750-9.

(208) US HUPO - US HUPO 2021 https://www.ushupo.org/Conference/ShortCoursestabid68Default.aspx (accessed Jun 1, 2020).

(209) Targeted proteomics: Experimental design and data analysis https://meetings.embo.org/event/19-proteomics (accessed Jun 1, 2020).

(210) Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, K.; Nesvizhskii, A. The Need for Guidelines in Publication of Peptide and Protein Identification Data: Working Group On Publication Guidelines For Peptide And Protein Identification Data. *Mol. Cell. Proteomics* **2004**, *3* (6), 531–533. https://doi.org/10.1074/mcp.T400006-MCP200.

(211) Hu, J.; Coombes, K. R.; Morris, J. S.; Baggerly, K. A. The Importance of Experimental

Design in Proteomic Mass Spectrometry Experiments: Some Cautionary Tales. *Brief. Funct. Genomic. Proteomic.* **2005**, *3* (4), 322–331. https://doi.org/10.1093/bfgp/3.4.322.
(212) Domon, B.; Aebersold, R. Options and Considerations When Selecting a Quantitative Proteomics Strategy. *Nat. Biotechnol.* **2010**, *28* (7), 710-.
(213) Weng, N.; Halls, T. D. J. Systematic Troubleshooting for LC/MS/MS. 19.
(214) Troubleshooting Liquid Chromatography-Tandem Mass Spectrometry in the Clinical Laboratory | AACC.org https://www.aacc.org/publications/cln/articles/2015/august/troubleshooting-liquid-chromatography-tandem-mass-spectrometry-in-the-clinical-laboratory (accessed Jun 2, 2020).
(215) Noga, M.; Sucharski, F.; Suder, P.; Silberring, J. A Practical Guide to Nano-LC Troubleshooting. *J. Sep. Sci.* **2007**, *30* (14), 2179–2189. https://doi.org/10.1002/jssc.200700225.
(216) Tabb, D. L. Quality Assessment for Clinical Proteomics. *Clin. Biochem.* **2013**, *46* (6), 411–420. https://doi.org/10.1016/j.clinbiochem.2012.12.003.
(217) Hubert, M.; Engelen, S. Robust PCA and Classification in Biosciences. *Bioinforma. Oxf. Engl.* **2004**, *20* (11), 1728–1736. https://doi.org/10.1093/bioinformatics/bth158.
(218) Rousseeuw, P. J.; Hubert, M. Robust Statistics for Outlier Detection. *WIREs Data Min. Knowl. Discov.* **2011**, *1* (1), 73–79. https://doi.org/10.1002/widm.2.
(219) Su, X.; Tsai, C.-L. Outlier Detection. *WIREs Data Min. Knowl. Discov.* **2011**, *1* (3), 261–268. https://doi.org/10.1002/widm.19.
(220) RISK6, a 6-gene transcriptomic signature of TB disease risk, diagnosis and treatment response | Scientific Reports https://www.nature.com/articles/s41598-020-65043-8 (accessed Jun 7, 2020).
(221) Chegou, N. N.; Black, G. F.; Kidd, M.; van Helden, P. D.; Walzl, G. Host Markers in Quantiferon Supernatants Differentiate Active TB from Latent TB Infection: Preliminary Report. *BMC Pulm. Med.* **2009**, *9* (1), 21. https://doi.org/10.1186/1471-2466-9-21.
(222) Thompson, E. G.; Du, Y.; Malherbe, S. T.; Shankar, S.; Braun, J.; Valvo, J.; Ronacher, K.; Tromp, G.; Tabb, D. L.; Alland, D.; Shenai, S.; Via, L. E.; Warwick, J.; Aderem, A.; Scriba, T. J.; Winter, J.; Walzl, G.; Zak, D. E.; Du Plessis, N.; Loxton, A. G.; Chegou, N. N.; Lee, M. Host Blood RNA Signatures Predict the Outcome of Tuberculosis Treatment. *Tuberculosis* **2017**, *107*, 48–58. https://doi.org/10.1016/j.tube.2017.08.004.
(223) Walzl, G.; McNerney, R.; du Plessis, N.; Bates, M.; McHugh, T. D.; Chegou, N. N.; Zumla, A. Tuberculosis: Advances and Challenges in Development of New Diagnostics and Biomarkers. *Lancet Infect. Dis.* **2018**, *18* (7), e199–e210. https://doi.org/10.1016/S1473-3099(18)30111-7.
(224) Henry, B. M.; Oliveira, M. H. S. de; Benoit, S.; Plebani, M.; Lippi, G. Hematologic, Biochemical and Immune Biomarker Abnormalities Associated with Severe Illness and Mortality in Coronavirus Disease 2019 (COVID-19): A Meta-Analysis. *Clin. Chem. Lab. Med. CCLM* **2020**, *1* (ahead-of-print). https://doi.org/10.1515/cclm-2020-0369.
(225) Ulhaq, Z. S.; Soraya, G. V. Interleukin-6 as a Potential Biomarker of COVID-19 Progression. *Med. Mal. Infect.* **2020**, *50* (4), 382–383. https://doi.org/10.1016/j.medmal.2020.04.002.
(226) Li, Y.; Wang, Y.; Liu, H.; Sun, W.; Ding, B.; Zhao, Y.; Chen, P.; Zhu, L.; Li, Z.; Li, N.; Chang, L.; Wang, H.; Bai, C.; Xu, P. *Urine Proteome of COVID-19 Patients*; preprint; Infectious Diseases (except HIV/AIDS), 2020. https://doi.org/10.1101/2020.05.02.20088666.
(227) *HUPO-PSI/MzQC*; HUPO Proteomics Standards Initiative, 2020.
(228) Yan, L.; Ma, C.; Wang, D.; Hu, Q.; Qin, M.; Conroy, J. M.; Sucheston, L. E.; Ambrosone, C. B.; Johnson, C. S.; Wang, J.; Liu, S. OSAT: A Tool for Sample-to-Batch Allocations in Genomics Experiments. *BMC Genomics* **2012**, *13* (1), 689. https://doi.org/10.1186/1471-

2164-13-689.
(229) Leek, J. T.; Scharpf, R. B.; Bravo, H. C.; Simcha, D.; Langmead, B.; Johnson, W. E.; Geman, D.; Baggerly, K.; Irizarry, R. A. Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data. *Nat. Rev. Genet.* **2010**, *11* (10), 733–739. https://doi.org/10.1038/nrg2825.
(230) Woolston, C. Potential Flaws in Genomics Paper Scrutinized on Twitter. *Nat. News* **2015**, *521* (7553), 397. https://doi.org/10.1038/521397f.
(231) Hu, A.; Noble, W. S.; Wolf-Yadlin, A. Technical Advances in Proteomics: New Developments in Data-Independent Acquisition. *F1000Research* **2016**, *5*. https://doi.org/10.12688/f1000research.7042.1.
(232) Bessant, C. *Proteome Informatics*; Royal Society of Chemistry, 2016.
(233) Ma, Z.-Q.; Dasari, S.; Chambers, M. C.; Litton, M. D.; Sobecki, S. M.; Zimmerman, L. J.; Halvey, P. J.; Schilling, B.; Drake, P. M.; Gibson, B. W.; Tabb, D. L. IDPicker 2.0: Improved Protein Assembly with High Discrimination Peptide Identification Filtering. *J. Proteome Res.* **2009**, *8* (8), 3872–3881. https://doi.org/10.1021/pr900360j.
(234) Sep 2018 – ACGT https://www.acgt.co.za/newsroom/2018/09/ (accessed Jun 11, 2020).
(235) 2018 SASBi/SAGS Conference – SAGS.
(236) HUPO-PSI meeting 2019 | HUPO Proteomics Standards Initiative http://www.psidev.info/hupo-psi-meeting-2019 (accessed Jun 11, 2020).
(237) Liu, F. T.; Ting, K. M.; Zhou, Z.-H. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*; 2008; pp 413–422. https://doi.org/10.1109/ICDM.2008.17.
(238) Thurstone, L. L. Multiple Factor Analysis. *Psychol. Rev.* **1931**, *38* (5), 406–427. https://doi.org/10.1037/h0069792.
(239) Schweppe, D. K.; Eng, J. K.; Bailey, D.; Rad, R.; Yu, Q.; Navarrete-Perea, J.; Huttlin, E. L.; Erickson, B. K.; Paulo, J. A.; Gygi, S. P. *Full-Featured, Real-Time Database Searching Platform Enables Fast and Accurate Multiplexed Quantitative Proteomics*; preprint; Cell Biology, 2019. https://doi.org/10.1101/668533.
(240) The National Laboratory Association https://www.home.nla.org.za/ (accessed Jun 11, 2020).
(241) Challenges and Opportunities for Biological Mass Spectrometry Core Facilities in the Developing World - PubMed https://pubmed.ncbi.nlm.nih.gov/29623005/ (accessed Jun 12, 2020).
(242) Home | SA-DIPLOMICS https://www.diplomics.org.za/ (accessed Jun 12, 2020).
(243) About Us – ACGT https://www.acgt.co.za/about-us/ (accessed Jun 12, 2020)

Addendum A:

Metric name	**Explanation**	**Units**	**MS1/ MS2**	**Formula if necessary**	**How to interpret:**
RT Divided metrics	The user has input a number of segments they want the data to be divided into, for example 10. The RT between the start of the run and the start+RTDuration is then divided by that number to obtain the upper boundary of the segments. Any scan with an RT lower than or equal to this scan is added to its segment. In our example there are 10 segments.				
MS2PeakWidths	For each peak within scans that are in the RT segment in question, the FWHM is calculated with Crawdad. This FWHM is then added to one array for the segment and the	Seconds	2		

	average of that array is reported.				
TailingFactor	Similar to peak widths, only this time the peak tailing factor is calculated by Crawdad thanks to modifications to the CrawDad source code as a submodule of Yamato.	Seconds	2	Calculated as stipulated in the USP31:(http://www.uspbpep.com/usp31/v31261/usp31nf26s1_c621.asp) W0.05 / (2 x f)	
MS2PeakCapacity	The size of each segment (Identical across segments) is divided by the average peak width of peaks in this segment.	Theoretical number of peaks that could be supported.	2	PeakCapacity is calculated as per Dolan et al.,2009, PubMed 10536823) Equation 1;	The larger the metric, theoretically the smaller the peak widths in that segment and the higher the number of peaks that can fit into the segment.

MS2PeakPrecision	For each basepeak, both the mz and intensity of every instance in which an ion (not necessarily a bpk) within the mztolerance of the basepeak is picked up is added to an array for mz and an array for intensity and the mean of those arrays are determined. The mean mz is divided by the mz at which the basepeak was reported as a basepeak. This value is squared and multiplied by the intensity mean. The intensity of the basepeak when it was reported as a basepeak is then divided by this result. Each peak is treated separately, despite the fact that the same m/z may be involved in a separate peak.	m/z	2	intensity(basepeak)/ mean(intensities of all occurrences)* (mean(mzs of all occurrences)/basepeakMz)^2	The larger the value, the less precise the m/z value of the peak. The value is weighted so as to not penalise low intensity peaks.
MS1PeakPrecision	Same as above, just for MS1	m/z	1	Same as above, just for ms1	
DeltaTICAvg	The TIC of a scan is subtracted from the TIC of the previous scan and the absolute value is then added to an array for all scans in the segment, of which the average is reported here...	Intensity	2		A larger value could indicate irregularities in the ionization process, such as sputter.
DeltaTICIQR	And the Interquartile range is reported here.	Intensity	2		A larger value could indicate a greater variation in the

					ionization process.
AvgCycleTime	If there are ms1scans: This measures the difference between the ms1 scan of a particular cycle and the starttime of the last ms2scan of that cycle. I know that we are then missing the time taking for the last scan of the cycle, however we do not have end times for scans with all instruments. So while we can use the starttime of the next cycle for most of the cycles as an end time, the very last cycle will not have this value and in a fixed window model, it will then look as though this scan is a different size to the others, when it is not necessarily. If there are no MS1 scans, the first ms2 scan of the cycle is used as starting point. This value is then added together for all cycles in a particular segment and the average is calculated.	seconds	1+2		A larger value could indicate a longer scan time which could be due to a larger amount of ions in this segment or other factors.

AvgMS2Density	For all the MS2 scans in a particular segment, the number of values in the number of ions detected (to be precise the number of values in the mz binary array) are added up and divided by the number of scans in that segment to get the average.	Counts	2		A larger value indicates a larger number of ions detected in this section of the RT on average.
AvgMS1Density	For all the MS1 scans in a particular segment, the number of values in the number of ions detected (to be precise the number of values in the mz binary array) are added up and divided by the number of scans in that segment to get the average.	Counts	1		A larger value indicates a larger number of ions detected in this section of the RT on average.
MS2TICTotal	For all the MS2 scans in a particular segment, the TIC values supplied by the mzML are added together.	Intensity	2		
MS1TICTotal	For all the MS1 scans in a particular segment, the TIC values supplied by the mzML are added together.	Intensity	1		
SWATHMetrics	For these metrics ms2scans are grouped by their isolationtargetwindows and those that are the same are				

	grouped into the same swath				
ScansPerSWATH	Number of scans for that same SWATH	count	2		
AvgMzRange	THis is the isolation window lower limit subtracted from the upper limit. This is also not technically a metric, but the user can use this information to make sense of the data.	m/z	2		
SwathProportionOfTotalTIC	The TIC value for all the scans that have the same isolation window target mz are added together. This value is then divided by the total TIC of all the MS2Scans.	proportion	2		
swDensityAvg	The number of ions detected (more precisely the number of values in the mz binary array) for all of the same swaths is added to an array and the mean of this array is reported.	count	2		
swDensityIQR	The value of Q3(75%ile) -Q1(25%ile) of above-mentioned array is reported.	count	2		
swAvgProportionSinglyCharged	Each time two m/z values in a scan is 1.00 +-0.001 apart we are making the assumption that they are M and M+1 peaks and that they are therefore singly charged. We count the number of these ions detected and divide that value by	proportion	2		

	the total number of ions in the scan as the proportion that are singly charged. We add these values for all the same swaths into an array and calculate the average of the array.				
Comprehensive metrics					
MissingScans	The number of scans where there was not a single ion detected.	count			
RTDuration	Difference between the first scan starttime and the last scan start time	minutes			
SwathSizeDifference	Difference between the largest swath and the smallest swath	m/z	2		
MS2Count	Number of MS2 scans in the entire run	count	2		
MS1Count	Number of MS1 scans in	count	1		

	the entire run				
SwathsPerCycle	Number of swaths in the same cycle	count	2		
TotalMS2IonCount	Number of ions detected in all MS2scans accross the run	count	2		
TotalMS1IonCount	Number of ions detected in all MS1scans accross the run	count	1		
MS2Density50	The median number of ions in all MS2 scans	count	2		
MS2DensityIQR	The IQR for the number of ions detected in all MS2 Scans	count	2		